Computational Studies of Docking and Amyloid Peptide Aggregation

Dissertation
zur Erlangung der naturwissenschaftlichen Doktorwürde
(Dr. sc. nat.)

Mathematisch-naturwissenschaftlichen Fakultät der
Universität Zürich

von
Marco Cecchini
aus
Italien

Promotionskomitee

Prof. Dr. Amedeo Caflisch
Prof. Dr. Markus Grütter

Zürich 2005

[1]Die vorliegende Arbeit wurde von der Mathematisch-naturwissenschaftlichen Fakultät der Universität Zürich auf Antrag von Prof. Dr. Amedeo Caflisch und Prof. Dr. Markus Grütter als Dissertation angenommen.

To Alessia who made it possible.
To her extraordinary strength and sincere love.

Summary

Proteins are complex, high molecular weight organic compounds involved in the most diverse and fundamental processes of the living organisms. To carry out their function, proteins have to fold into a unique three-dimensional structure, called the "native" state, that is exclusively determined by the amino acid sequence. The "native" state is the translation of the genetic information and represents the *meaning* of the sequence in the language of proteins. The statistical mechanic view of protein folding describes the folding reaction as a diffusion of an ensemble of polypeptide chains on a funnel-like energy landscape. Within this framework, the complete description of the free-energy landscape of individual proteins represents a synthetic way to describe both thermodynamic and kinetic aspects of protein folding and opens the way toward the breaking of the "protein code". Despite continuous development of innovative experimental techniques, the determination of complete free-energy surfaces of individual proteins is not yet feasible. Only the synergy between experimental and computational strategies can supply information at the desired level of detail. Computational approaches to simulate the behavior of model proteins at atomic resolution (*in silico* experiments) can be invoked. In the present thesis two *in silico* experiments are presented: molecular docking for drug discovery and amyloid peptide aggregation.

In molecular docking, the "virtual experiment" aims at the identification of molecules able to bind the key regions of pharmacologically relevant enzymes or macromolecules with high affinity and selectivity. The computational approach mimics the molecular recognition between the target receptor and small organic compounds on a computer by means of a simplified energy model and a searching procedure. In this thesis, an improved version of the fragment-based flexible ligand docking approach SEED-FFLD developed in house is presented. The implementation of a "hybrid procedure" to search the conformational space of the ligands significantly improves the quality of the SEED-FFLD docking predictions at a moderate additional computational cost. The docking strategy is tested on highly flexible inhibitors of the human immunodeficiency virus type 1 protease, human α-thrombin and the estrogen receptor β. The docking results indicate that it is possible to correctly reproduce the binding mode of inhibitors with more than ten rotatable bonds if the strain in their covalent geometry (i.e., its bond angles and lengths) upon binding is not large. Hence, for docking a limited set of compounds, strategies that allow for full flexibility of the ligand, though computationally very expensive, should be preferred. Thanks to the methodological development, the in-house docking procedure has been recently applied to screen large libraries of compounds against β-secretase, a very difficult target involved in Alzheimer's disease. Remarkably, low-micromolar activity was measured *in vitro* for several compounds

suggested by the docking procedure.

In the study of amyloid peptide aggregation, the ultimate goal of the *in silico* experiment is to simulate fibrilization within a sample of amyloid proteins and determine the complete free-energy surface for the oligomeric system. The atomic detail of the simulation could provide a complete structural description of the molecular intermediates along fibrilization, which are believed to be the most cytotoxic species. *In vitro* fibril formation is a rather slow process and dependently on amyloid protein sequence may take from several minutes to days. Such timescales are never accessible by standard computer simulations. Moreover, amyloidogenesis is intrinsically cooperative and therefore several aggregating units must be included in the model, thus dramatically increasing the possibility of statistical errors. In the present thesis, the simulation results show that it is possible to investigate the early steps of amyloid aggregation at physiologically relevant temperature values by using efficient sampling procedures and simplified energy models, such as the replica exchange molecular dynamics (REMD) approach and an implicit treatment of the solvent. In agreement with experimental evidence, it has been shown that the nucleation process, i.e., the formation of amyloid nuclei, can be interpreted as a condensation stage toward disordered aggregates followed by an order transition. Therefore, the simulation results were analyzed with two order parameters borrowed from liquid crystal theory. Interestingly, it has been observed that the nematic order parameter $\overline{P_2}$ averaged over the canonical ensemble effectively estimates the β-aggregation propensity of a peptide system and discriminates amyloidogenic from soluble peptides in agreement with the experiments. From this operational definition of β-aggregation propensity, a novel *in silico* approach to investigate the aggregation properties of amyloid polypeptides has been derived. A novel strategy has been also designed to predict the position dependence of the β-aggregation propensity along the protein sequence, thus highlighting possible amyloidogenic stretches (the aggregation *hot-spots*). The β-aggregation propensity along the sequence of the Alzheimer's amyloid-β peptide has been investigated by multiple molecular dynamics simulations of oligomeric systems of 7- and 11-residue segments for a total of 0.31 milliseconds. The β-aggregation propensity is found to be highly heterogeneous with a maximum in the segment $\mathtt{V}_{12}\mathtt{HHQKLVFFAE}_{22}$ and minima at $\mathtt{S}_{8}\mathtt{G}_{9}$, $\mathtt{G}_{25}\mathtt{S}_{26}$, $\mathtt{G}_{29}\mathtt{A}_{30}$, and $\mathtt{G}_{38}\mathtt{V}_{39}$ which are turn-like segments. Similar findings are obtained for the human amylin, a 37-residue peptide which displays a maximal β-aggregation propensity at $\mathtt{Q}_{10}\mathtt{RLANFLVHSSNN}_{22}$ and two turn-like sites at $\mathtt{G}_{24}\mathtt{A}_{25}$ and $\mathtt{G}_{33}\mathtt{S}_{34}$. In the last application, the MD approach is used to identify β-aggregation *hot-spots* within the N-terminal domain of the yeast prion Ure2p ($\mathrm{Ure2p}_{1-94}$) and to design a double-point mutant ($\mathrm{Ure2p\text{-}N4748S}_{1-94}$) with lower β-aggregation propensity.

Zusammenfassung

Proteine sind komplexe organische Verbindungen von hohem Molekulargewicht, die an den unterschiedlichsten und fundamentalsten Prozessen lebender Organismen beteiligt sind. Um ihre Funktion auszuüben, müssen sich Proteine in eine eindeutige dreidimensionale Struktur falten, den sogenannten nativen Zustand, welcher einzig und allein durch die Aminosäuresequenz eines Proteins bestimmt ist. Der native Zustand ist die Übersetzung der genetischen Information und entspricht der Bedeutung der Sequenz in der Sprache der Proteine. Aus Sicht der statistischen Mechanik entspricht die Faltungsreaktion einer Diffusion eines Ensembles von Polypeptidketten auf einer trichterähnlichen Energielandschaft. In diesem Zusammenhang stellt die vollständige Beschreibung der freien Energielandschaft eines Proteins eine effektive Art dar, die thermodynamischen und kinetischen Aspekte der Proteinfaltungsreaktion zu beschreiben und öffnet den Weg zum Knacken des "Proteincodes". Trotz stetiger Entwicklung von innovativen experimentellen Techniken kann die vollständige freie Energielandschaft zur Zeit für kein einziges Protein bestimmt werden. Einzig die Synergie zwischen experimentellen und computergestützten Strategien kann die Information im gewünschten Detail liefern. Computergestützte Simulationen des Verhaltens von Modellproteinen in atomarer Auflösung können dazu herangezogen werden (*in silico* Experimente). In dieser Doktorarbeit werden zwei solcher *in silico* Experimente vorgestellt: Molekulares Docken zur Wirkstoffentwicklung und die Aggregation von Amyloidpeptiden.

Im ersten Fall ist das Ziel der "virtuellen" Experimente die Identifizierung von Molekülen, die mit hoher Affinität und Selektivität an Schlüsselregionen von pharmakologisch relevanten Makromolekülen binden. Der computergestützte Ansatz stellt am Computer die molekulare Erkennung zwischen anvisiertem Rezeptor und kleinen organischen Molekülen durch die Verwendung von vereinfachten Energiemodellen und eines Suchverfahrens nach. In dieser Doktorarbeit wird eine verbesserte Version des auf Fragmente basierenden Ansatzes zum Docken von flexiblen Liganden namens SEED-FFLD vorgestellt, welcher in unserer Forschungsgruppe entwickelt wurde. Die Implementierung einer Hybridprozedur aus lokaler Suche und genetischem Algorithmus zum Absuchen des Konformationsraumes eines Liganden, verbessert die Qualität der SEED-FFLD Vorhersagen deutlich bei nur begrenzter Rechenzeitzunahme. Diese Strategie zum Docken ist an extrem flexiblen Inhibitoren der Protease des menschlichen Immunschwäche-Virus Typ 1, des menschlichen α-Thrombins und des Östrogenrezeptors β getestet worden. Die Resultate weisen darauf hin, dass es möglich ist, den Bindungsmodus von Inhibitoren mit mehr als zehn rotierbaren Bindungen korrekt zu reproduzieren, falls die intramolekulare Spannung des Inhibitors beim Binden nicht zu stark zunimmt. Aus diesem Grund sollten für das Docken von einer begrenzten Zahl von Verbindungen Strategien

bevorzugt werden, die eine vollständige Flexibilität der Liganden voraussetzen, ungeachtet der Rechenzeitzunahme. Dank der Weiterentwicklung der Methode konnte unsere Prozedur zum Docken vor Kurzem benutzt werden, um grosse Bibliotheken von Verbindungen nach möglichst guten Inhibitoren der β-Sekretase durchzusuchen - ein Enzym, das in der Alzheimerschen Krankheit eine wichtige Rolle spielt und eine grosse Herausforderung an unsere Wirkstoffentwicklungstrategie darstellt. Bemerkenswerterweise wurde für drei der von unserer Dock-Prozedur vorgeschlagenen Verbindungen eine *in vivo* Aktivität im unteren mikromolaren Bereich gemessen.

Das Ziel des zweiten *in silico* Experiments ist es, die Aggregation eines Ensembles von Amyloidproteinen zu simulieren und die vollständige freie Energielandschaft dieses Systems zu bestimmen. Die atomare Auflösung dieser Simulationen ermöglicht eine vollständige Beschreibung der Struktur der während der Bildung von Fibrillen auftretenden molekularen Zwischenprodukte, welche für die am meisten zytotoxischen Spezies gehalten werden. Die *in vitro* Bildung von Fibrillen ist ein langsamer Prozess, der, je nach Amyloidproteinsequenz, von einigen Minuten bis hin zu ganzen Tagen dauern kann. Solche Zeitskalen liegen in keiner Weise im Bereich von gewöhnlichen Computersimulationen. Ausserdem ist die Amyloidogenese ein intrinsisch kooperativer Prozess und das Modell muss daher einige aggregierende Einheiten enthalten, was jedoch die Wahrscheinlichkeit vom Auftreten von statistischen Fehlern dramatisch erhöht. Die Simulationsresultate, die in dieser Doktorarbeit vorgestellt werden, zeigen, dass es möglich ist, *in silico* die ersten Schritte der Amyloidaggregation zu untersuchen, wenn effiziente "sampling"-Methoden und vereinfachte Energiemodelle, wie z.B. der sogenannte "replica exchange molecular dynamics"-Ansatz und die implizite Behandlung des Lösungsmittels, verwendet werden. Im Einklang mit experimentellen Daten konnte gezeigt werden, dass der Nukleationsprozess, d.h. die Bildung des Amyloidkerns, als Kondensation zu ungeordneten Aggregaten interpretiert werden kann, welche von einem Ordnungsübergang gefolgt wird. Die Simulationsresultate wurden daher mittels zweier Parameter analysiert, die der Flüssigkristalltheorie entlehnt worden sind. Interessanterweise wurde beobachtet, dass der Mittelwert des nematischen Ordnungsparameters $\overline{P_2}$ über das kanonische Ensemble die Tendenz von Peptidsystemen β-Aggregate zu bilden auf effektive Art und Weise abzuschätzen vermag und, im Einklang mit experimentellen Daten, amyloidogene von löslichen Peptiden unterscheiden kann. Ausgehend von der *in silico* Definition der β-Aggregationstendenz wurde ein neuer Ansatz zur Untersuchung der Aggregationseigenschaften von amyloidogenen Proteinen entwickelt. Die Strategie wurde so ausgelegt, dass Vorhersagen über die Positionsabhängigkeit der β-Aggregationstendenz entlang der Proteinsequenz möglich sind und so eventuell vorhandene amyloidogene Segmente ausfinding gemacht werden können. Die β-Aggregationstendenz entlang der Sequenz des Alzheimerschen Amyloid β-Peptids ist ($A\beta_{42}$) in mehreren

Moleküldynamiksimulationen - insgesamt 0.31ms Simulationszeit - von oligomeren Systemen von 7 und 11 Reste langen Segmenten untersucht worden. Die β-Aggregationstendenz ist sehr heterogen mit einem Maximum im Segment $\mathtt{V}_{12}\mathtt{HHQKLVFFAE}_{22}$ und mehreren Minima an den Positionen $\mathtt{S}_{8}\mathtt{G}_{9}$, $\mathtt{G}_{25}\mathtt{S}_{26}$, $\mathtt{G}_{29}\mathtt{A}_{30}$, und $\mathtt{G}_{38}\mathtt{V}_{39}$. Letztere entsprechen schleifenähnlichen Segmenten. Ähnliche Befunde werden für das menschliche Amylin beobachtet. Amylin ist ein Peptid bestehen aus 37 Aminosäuren, das eine maximale β-Aggregationstendenz für das Segment $\mathtt{Q}_{10}\mathtt{RLANFLVHSSNN}_{22}$ und die zwei schleifenähnlichen Positionen $\mathtt{G}_{24}\mathtt{A}_{25}$ und $\mathtt{G}_{33}\mathtt{S}_{34}$ aufweist. Als letzte Anwendung wurde der Moleküldynamikansatz verwendet, um sogenannte *hot spots* für die β-Aggregation in der N-terminalen Domäne des Hefe-Prions Ure2p ($\mathrm{Ure2p}_{1-94}$) zu identifizieren, und um die Doppelmutante ($\mathrm{Ure2p\text{-}N4748S}_{1-94}$) mit einer geringeren β-Aggregationstendenz zu generieren.

Contents

Chapter 1

Introduction

1.1 Protein Molecules

Proteins (in Greek $\pi\rho\omega\tau\epsilon\nu\eta$ = *first element*) are complex, high molecular weight organic compounds essential to the structure and function of living organisms. Chemically speaking, they are linear polymers of the same type, built of various combinations of 20 building blocks, i.e., the natural amino acids, and they differ only in the sequence in which the blocks are assembled and the length of the polymeric chain. Hence, proteins constitute a relatively homogeneous class of molecules. Despite the apparent simplicity of their chemical structure, proteins are involved in the most diverse and fundamental biological processes. Some of them, called *enzymes*, act as biological catalyzers and permit the occurrence of essential chemical reactions under physiological conditions (i.e., aqueous solution, 37°C, *p*H 7, atmospheric pressure). Some others store and transport a variety of particles ranging from macromolecules to electrons, transmit information between specific cells and organs, control the passage of molecules across the cellular compartments, defend the organism against intruders (*antibodies*), convert chemical energy into mechanical energy and control gene expression. Thus, if we aspire to understand how living organisms function, then we must first understand the behavior of proteins. In order to carry out their biological function, proteins have to fold into a unique three-dimensional structure (called the "native" state) which is exclusively determined by the amino acid sequence [1]. The "native" state of a protein is the translation of the genetic information encoded in the sequence and from a semantic point of view, it represents the *meaning* of the sequence in the language of proteins. In such a fascinating language, the 20 amino acids are the letters of the alphabet and the naturally occurring sequences "meaningful" words and sentences. The semantic interpretation of the amino acid sequence is rather intuitive and provides a simple description of most of the known properties of proteins (see Fig. 1.1). As in any language, for example, only certain combinations of letters are meaningful whereas random or incomplete strings often convey

gibberish (a). Words and sentences have a sequence directionality which allows them to be decoded (b) and they incorporate a certain robustness that prevents them from being misunderstood even in case of accidental errors (c). Moreover, sharing common portions (substrings) strings and sentences can bring completely different messages (d). A few, instead, are degenerate and potentially own multiple meanings which depend on the local context (e). Finally, diverse strings with non detectable sequence identity share similar significance and can be alternatively used in the same context (f).

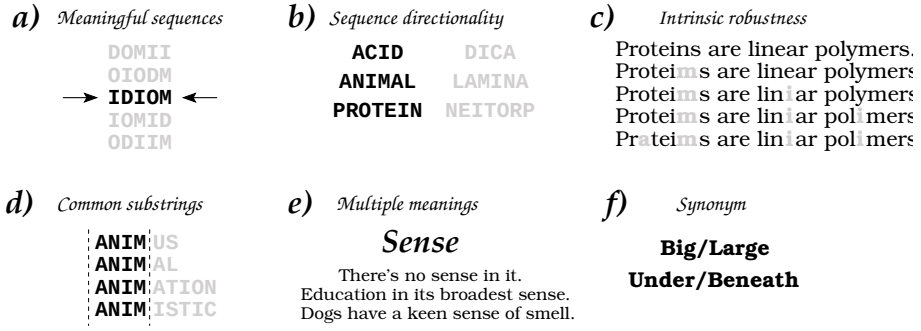


Fig. 1.1: The semantic interpretation of protein sequences

All these properties are common to protein sequences. The latest release of the TrEMBL database (release 46.5, April 2005) contains 1,662,660 ($\sim 10^6$) unique protein sequences. Considering an average length of 200 residues per sequence, the total number of possible amino acid combinations is 200^{20} ($\sim 10^{46}$). Despite the rough estimate, this difference of 40 orders of magnitude indicates that naturally occurring sequences are a tiny minority (a). Retro protein-folding experiments have revealed that retro proteins are no more similar to their parent sequences than random sequences, despite having the same amino acid composition [2] (b). The biological function of a protein has been shown to be more correlated to the macromolecular geometry than to chemical detail [3]: when the global three-dimensional structure and the active site of a protein are conserved, considerable modifications of the sequence can be made without any loss of function (c). Despite sharing short to medium sized subsequences, many of these proteins do not present common folds nor do they accomplish similar tasks (d). Certain sequences, like those from *amyloidogenic* proteins such as transthyretin, the prion protein (PrP) and the human lysozyme, code for folded structures with a high conformational plasticity that allows them to refold into a stable "non-native" conformation under physiological conditions. The information encoded is therefore degenerate and corresponds to different "meanings" depending on the local context (e). Finally, protein

structures are much more conserved than protein sequences [4]. Therefore, sequences with low or non-detectable similarity can still promote the same fold (f).

The semantic interpretation of the amino acid sequence suggests that proteins are encoded in a proper language, the interpretation of which would lead to an outstanding result: the solution of the mysteries of life and evolution. If the language could be deciphered, in fact, the still missing link between "protein sequence" and "specific function" would be discovered, thus opening the way towards the understanding of biological processes at atomic level. Molecular details of abnormal processes associated with severe pathologies could be finally unveiled and highly specific and bio-compatible molecules could be designed to deal with them. The breaking of the "protein code" would promote the development of efficient strategies for treating and preventing human disorders by means of *drugs* able to speak the language of the pathogens. Unfortunately, the decoding process is far from trivial and although the ultimate goal is very appealing, the "protein language" has not yet been deciphered.

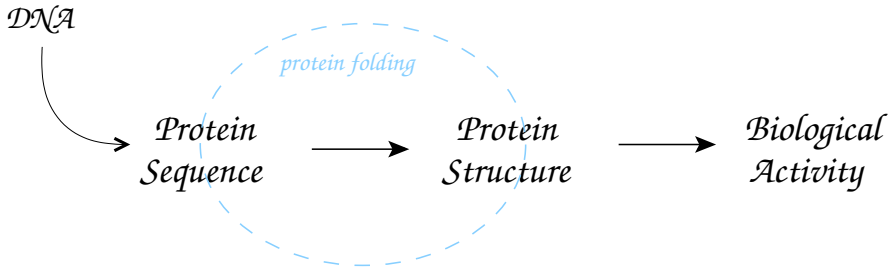


Fig. 1.2: From protein sequence to protein function

As proteins must fold to fulfill their biological function, the relation between the amino acid sequence and the biological activity is clearly mediated by the "native state", the structure designed by evolution for the particular task. Hence, to *speak* the language of proteins, i.e., to be able to develop effective therapeutic strategies for treating human diseases, one must first understand the link between protein sequences and protein structures: the "protein folding" problem (see Fig. 1.2).

Protein Folding

Under physiological conditions, proteins fold to the native state rapidly and reliably. To this purpose, protein sequences must satisfy two requirements: one "kinetic" and one "thermodynamic". The *thermodynamic requirement* is that the three-dimensional structure of a protein in its physiological milieu (solvent, pH, ionic strength, presence of other components such as metal

ions or prosthetic groups, temperature, etc.) corresponds to the global minimum of the Gibbs free energy. The *kinetic requirement* is that the denatured polypeptide chain folds into the native conformation in a reasonable amount of time. Given the huge number of degrees of freedom involved, satisfying the kinetic requirement is not trivial. Even assuming only three permissible configurations per residue, a medium sized chain of 100 amino acids would be able to assume 3^{100} ($\sim 10^{47}$) different conformations in solution. If only 10^{-11}s were required to convert from one conformation into another, the complete search of the conformational space would require about 10^{28} years [5]. Nevertheless, most proteins reach their native state in the timescale of seconds or minutes. This puzzling inconsistency, better known as the "Levinthal's paradox", has slowed down the understanding of protein folding until a new era of experiments has shed light on the early events of the process at atomic resolution. Thanks to great advances in both experiments and theory, the classical "macroscopic" view of protein folding based on simple phenomenological models was substituted by a "new view" that describes folding as a statistical mechanics process. Macroscopic states of the folding reaction (the folded, unfolded and intermediate states) have been interpreted in terms of *ensembles* of individual conformations (microstates) and the pathway concept of sequential events have been replaced by the funnel concept of parallel events. In the new view, the folding process is not described by the trajectory of a single molecule on a shallow surface (Fig. 1.3, b) but by the diffusion of an ensemble of asynchronous chains on a funnel-like energy landscape (Fig. 1.3, c). Or quoting K.A. Dill:

> *"During folding, the individual chains move on the funnel surface and ultimately find their ways to the same native structure in the same way that water flowing along different routes down a mountainside can ultimately reach the same lake at the bottom."*

The new view of protein folding uses the concept of "free-energy surfaces" and "folding funnels" to describe the folding reaction. These concepts capture the deep essence of the problem and interpret the universality of the folding mechanism despite substantial differences among proteins. In particular, the interpretation of folding via free-energy surfaces provides a concise and useful framework for studying such complex systems. Considering protein folding as a chemical reaction, the free-energy surface is the analogue to the potential energy surface (PES) for simple chemical reactions [7]. In the latter case, the PES describes the energy of interaction of the atoms involved as a function of their position and provides the complete description of both thermodynamics and kinetics of the reaction. A detailed knowledge of the PES and the laws of dynamics enable the calculation of the trajectories along which molecules move from reactants to products, the identification of the transition state and the intermediates along the reaction pathways and the determination of the overall rate constant [8, 9]. By analogy, a complete

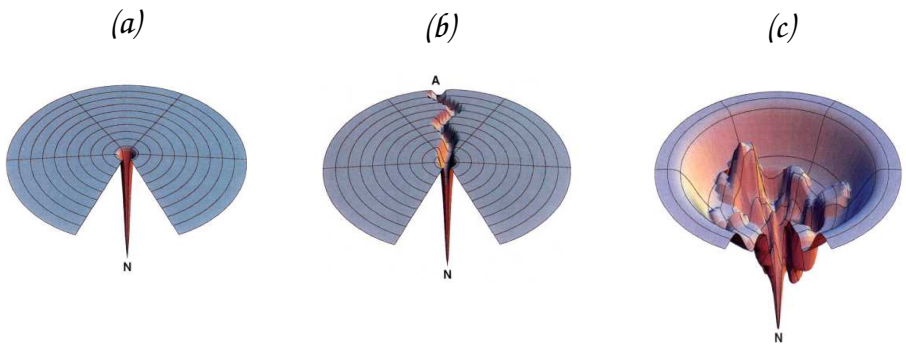


Fig. 1.3: Energy landscapes for protein folding: the Levinthal "golf-course" (a), the "pathway" solution to the search problem (b) and the rugged funnel with kinetic traps, barriers and narrow paths to the native state (c). N is the native conformation. (Adapted from Dill and Sun Chan [6].)

description of the free-energy surface of a protein can provide the same information for the folding reaction. However, protein folding is a complex process which involves a large number of degrees of freedom. Thus, the relative free energy surface has a multidimensional character and includes a multitude of local minima and transition regions. To obtain a meaningful description of the accessible states and energetic barriers of the system, a suitably defined progress variable has to be chosen. The projection of the free-energy on to the progress variable, the value of which is obtained by averaging over non-essential degrees of freedom, provides a simplified but rather intuitive view of the mechanism which still accounts for the complexity of the process, i.e., the existence of intermediate species "on" and "off" the folding pathway, with the latter responsible for the heterogeneity of the folding rates.

Indeed, free-energy surfaces are extremely valuable and their complete description provides a unique tool for understanding the determinants of many aspects of protein behavior including the stability of the native state, the kinetics for folding and the mechanism of misfolding. They represent a quantitative link between protein sequence and biological activity and help in decoding the "language" of proteins. Experimentally, the complete determination of free-energy surfaces for individual proteins is not feasible yet, despite the continuous development of innovative techniques. Nowadays, only the synergy between experimental and computational strategies, here referred to as *in silico* experiments, can provide the description of free-energy landscapes at the desired detail. The relevance of computational approaches to the comprehension of the language of proteins and their high complementarity with experiments are discussed in the next section.

Whilst waiting for further progress in the interpretation of the protein language, effective therapeutic strategies for treating human diseases can be developed by focusing on the specific activity of proteins. Once a pharmacologically relevant target has been identified, small molecules, that are able to bind to its key or "active" regions and to hinder its biological function, can be researched. This alternative approach does not require any knowledge of the folding mechanism and has proved very effective especially when the three-dimensional structure of the target is available. In this field, the synergy between experimental and computational strategies is very well established and even "only-for-profit" pharmaceutical companies routinely follow *mixed* procedures to discover new effective drugs. In particular, when the crystal structure of the target is known, *structure-based in silico* experiments can be designed to mimic the protein-ligand molecular recognition patterns and estimate the binding affinities of electronic libraries of compounds in a very fast and cheap way. A basic framework for computational procedures used for "drug discovery" is also presented in the next sections.

1.2 *In silico* Experiments

Experimental Background

Thanks to recent advances in experimental techniques, substantial progress has been made towards the understanding of the mechanisms lying behind the structure and function of proteins. In particular, the determination of three-dimensional structures of proteins and protein complexes by X-ray crystallography and NMR spectroscopy has provided essential insights, i.e., the location of the active site, information on the catalytic mechanism and the possible conformational changes. Examples of recent spectacular achievements include the determination of the structure of F_1-ATPase [10], the proteasome [11] and the ribosome [12, 13]. Although the total number of solved structures (31,059 entries currently stored in the PDB database, May 2005) is continuously increasing, this number represents "only" about 2% of all known protein sequences. Very accurate X-ray and NMR experiments are, in fact, time consuming, not always successful and require a lot of human expertise. In addition, their applicability is limited to some extent. Moreover, both methods provide "only" a static picture of the problem, thus neglecting not only important aspects of protein behavior involving non-native structures [14] but also internal motions of the protein that are crucial for the biological activity. Indeed, to capture these aspects, an accurate description of the overall protein dynamics is required.

To this purpose, novel applications of established experimental techniques have been tried to monitor the structural rearrangements of a polypeptide chain. However, it is never straightforward to obtain information at atomic detail. To increase data resolution, mixed techniques have been pro-

posed, which monitor different structural aspects at the same time. A good example is the combined usage of far- and near-UV circular dichroism (CD): while the former approach determines the average content of the secondary structure, the latter monitors the packing of aromatic side-chains. *Mixed* methods have provided a more global, though not yet complete, description of protein conformational changes during folding [15].

The time resolution of the experiment is another problematic aspect. Approaches limited to events on the millisecond range or longer, such as standard CD, are not suitable to monitor rapid structural changes occurring on nano to microsecond timescales. The development of novel methods based on fluorescence and infrared sprectoscopy (IR) have extended the range of available time-windows and successfully probed the early events in protein folding [16, 17].

In a folding reaction, a highly heterogeneous ensemble of molecules, that differ substantially in their structure and dynamics, is involved. For such systems, characterized by a stochastic behavior, it is impossible to detect the dynamics of individual molecules by ensemble-averaged measurements. Therefore, standard experimental techniques cannot provide accurate information concerning molecular states (local minima of the free-energy surface), especially far away from the native one. To overcome these difficulties, novel techniques called SMD methods (single-molecule detection), have recently been proposed [18, 19, 20]. SMD techniques, among which atomic force microscopy (AFM), total internal reflection fluorescence microscopy (TIRFM), optical-trapping nanometry, polarized fluorescence and fluorescence resonance energy transfer (FRET) are the most popular, can monitor the time evolution of single biomolecules during their functional activity and allow the detection of global movements and conformational changes. SMD have been successfully applied to study the dynamic properties of motor proteins [21], enzymes [22], RNA-polymerase [23] and cell-signaling proteins [24, 25]. However, the rather low resolution of these experiments is still a stringent limitation.

More detailed analysis of fundamental processes such as protein folding and molecular docking requires monitoring the nature and the energies of the interactions between individual atoms as a function of time. NMR spectroscopy and site-directed mutagenesis can partly supply this kind of information. In the former approach, the proximity of specific pairs of atoms can be detected through the measurement of nuclear Overhauser effects (NOEs). The ensemble of conformations that satisfy the experimental distances is then determined. In the latter approach, the properties of the wild-type protein are compared with those of a series of mutants [26]. In summary, it is assumed that a residue is involved in native contacts in the transition state if a reduction in the size of its side chain upon alanine mutation destabilizes the transition state as much as it destabilizes the native state. Within the validity of this assumption, the method provides a structural description of

the transition state ensemble (TSE).

Despite great progress, current experimental strategies are not sufficient to provide a "complete" description of protein free-energy landscapes. As pointed out, some techniques are too focused on the native state structure and do not provide any information about the dynamics; others present time or space resolution problems and cannot properly describe the time evolution of the polypeptide chain, i.e., protein internal motions and conformational changes. Other methods supply information based on averages over heterogeneous ensembles and thus lose much of the static and dynamic detail.

Computational Approach

In an attempt to overcome the limitations described above, an alternative approach would be to perform an experiment *in silico*: studying the behavior of a model protein by means of a computational protocol. In principle, by using an accurate atom-based model for the potential energy (a force field) and solving the time-discretized form of Newton's equation of motion in the presence of the appropriate solvent, it should be possible to reproduce the exact dynamics of a polypeptide chain and monitor the complete folding trajectory at atomic resolution. If this were feasible, *in silico* experiments could be used to address specific questions about proteins much more easily than the experiments themselves. However, computer simulations have not reached this stage (yet) and, similarly to experimental techniques, they suffer from stringent limitations. For a model protein of about 100 residues, the complete folding transition *in vitro* takes about 1 ms. With the available simulation protocols and computing power, monitoring the folding trajectories of such a protein would require more than 10 years. Moreover, current force fields even in their most sophisticated forms are not accurate enough to let a protein fold on a computer. Even if one could use a computer several orders of magnitude faster than currently available processors, thus solving the timescale problem, the model protein would never find its way to the native state. *Statistical errors* related to timescales and sampling of configuration space and *systematic errors* related to the inaccuracy of the energy models dramatically reduce the applicability of such approaches. Nonetheless, computer simulations can provide the ultimate details concerning the motion of individual atoms as a function of time, which will never be accessible from a sample in a test tube. In their range of validity, *in silico* experiments are a valuable complement to experimental approaches. On the other hand, experiments play a crucial role in defining that range of validity: comparisons between simulation and experiments estimate the reliability of the *in silico* results and provide useful criteria for improving the simulation methodology. Hence, *in vitro* and *in silico* experiments should be considered as complementary side views of the same problem and not mutually exclu-

sive approaches. In the following, a series of notable examples highlight the potential synergy between the two diverse approaches.

Role of solvent in protein dynamics: A particularly striking example is provided by the resolution of the contentious question concerning the role of solvent on the internal motion of proteins [27]. Experimentally, it has not been possible to determine whether solvent fluctuations drive the internal motion of proteins. This question was addressed by a computational study recreating a physical system which is not accessible in Nature. Molecular dynamics simulations were performed with one part of the system (the protein) at one temperature and the other part (the solvent) at a different temperature. Protein atomic fluctuations calculated from the simulation trajectories at either 180 or 330 K revealed that their magnitude is only weakly dependent on protein temperature. In contrast, the fluctuations are large when the solvent is at 300 K and small when the solvent is at 180 K, independent of the protein temperature. This result demonstrates that the temperature of the solvent, and thus its mobility, is the dominant factor in determining the functionally important protein fluctuations under physiological conditions.

Conformational change in the functional mechanism of GroEL: Another famous *in silico* experiment have been designed to find a pathway between the open and closed conformations of the bacterial chaperonin GroEL, which has been impossible to determine experimentally. GroEL is supposed to assist the folding of about 10% of cytosolic proteins in *Escherichia coli*. During its functional cycle, GroEL undergoes large conformational changes [28] that are regulated by the binding and hydrolysis of ATP and involve the co-chaperonin GroES. In the absence of ATP and GroES, the seven subunits of GroEL are in a "closed" conformation; in contrast, they adopt an "open" conformation when bound to the cofactors. Since the actual transition between the "open" and "closed" conformations occurs in the millisecond timescale, a direct simulation of the overall motion by molecular dynamics would not be possible. However, several methods have been developed to follow the transition between two experimentally determined states on shorter timescales. One of these, targeted molecular dynamics (TMD), was actually applied to determine the transition pathways of GroEL [29]. The computational results indicated that, in the absence of GroES, the subunits adopt an intermediate conformation upon ATP binding. The simulations unveiled that the motion of the intermediate domain induced by nucleotide binding triggers the larger movement of the apical and equatorial domains. Subsequent cryoelectron microscopy results have confirmed this prediction.

Protein Folding: *In silico* experiments have been successfully applied to

study the reversible folding of structured peptides [30, 31, 32, 33]. Suitably designed molecular dynamics simulations have reproduced the correct folding of a three-stranded β-sheet (beta3s) [31] and displayed multiple folding pathways at atomic resolution [34]. The information provided by the simulations, including the complete determination of the free energy landscape and the description of the folding mechanism, could never be obtained experimentally. Since the fastest folding reactions require $\sim 10\mu$s to complete, folding a protein on a computer using brute force techniques is extremely difficult (mainly due to statistical errors). Simplified models and effective protocols that allow sampling at longer timescales are then required. A common approach for protein folding is to treat water molecules "implicitly" by replacing them with a continuum dielectric. In this way, implicit solvation models eliminate the solvent degrees of freedom and allow individual simulation runs to explore the microsecond timescale. For beta3s an implicit model based on the solvent accessible surface area [35] was applied. Despite the relative "inaccuracy" of the energy model, a statistically relevant analysis of the folding process of this 20-residue peptide could be performed. So far, this kind of approach has only been successful with short peptides. Despite the limited biological relevance, these systems are useful models to investigate the general aspects of protein folding.

Structure determination of protein non-native states: Characterizing the nature of partially folded states is crucial for understanding the thermodynamic and kinetic behavior of protein molecules. Experimental approaches used for protein determination involve three major steps: (i) the choice of a technique to obtain data that can be interpreted in terms of structural parameters, i.e., atomic distances, dihedral angles, solvent exposure, etc.; (ii) the selection of an appropriate energy model to represent the structure and the energy of the molecule; (iii) the definition of an optimization technique to select conformations that minimize the deviations from experimental data. This approach has proved extremely successful for the description of the native state of protein (by X-ray or NMR) but rather inadequate for the characterization of non-native states that are only partially or transiently structured (unfolded state or TSE). In the latter case, the main limitations arise from the heterogeneity of the ensembles of conformations involved. A novel and successful approach to characterize these elusive non-native ensembles combines a computational strategy with experimental measurements that are used to restrain the conformational space. These *in silico* experiments aim to find all molecular conformations that minimize a pseudo-energy function composed of two parts: the first defines the physico-chemical properties of the polypeptide chain in its environment; the second penalizes deviations from experimentally derived parameters [36, 37]. Biased MD simulations ensure, on one hand, that computer generated models are meaningful and, on the other hand, that they are compatible with experimental measurements.

A striking example of the described protocol is given by the structure determination of the TSE for acylphosphatase (AcP) by using experimental ϕ-values [36]. Assuming that ϕ-values can be approximated by the fraction of native interactions formed in the transition state, one can work out a simple additional energy term to bias the conformational search:

$$\rho = \frac{1}{N_\phi} \sum_{i \in E} (\phi_i - \phi_i^{exp})^2 \ , \tag{1.1}$$

where E is the list of the N_ϕ available experimental ϕ-values, ϕ_i^{exp}. The ϕ-value of amino acid i in the conformation at time t is defined as:

$$\phi_i(t) = \frac{N_i(t)}{N_i^{\mathrm{nat}}} \ , \tag{1.2}$$

where $N_i(t)$ is the number of native contacts of residue i at time t and N_i^{nat} the number in the native state. The procedure lets TSE conformations become the most stable states on the potential energy surface, thus allowing statistically relevant sampling and structure determination of TSE. Interestingly, for AcP it has been shown that using just three specific ϕ-values, all the remaining ϕ-values could be predicted with high precision, thus suggesting that the ensemble of conformations determined by the computational approach is a faithful representation of the TSE.

Paradigm of the *in silico* Experiments

The reported examples clearly illustrate the relevance of computational strategies in structural biology and underline the potential complementarity with *in vitro* experiments. Different protocols have been successfully proposed to provide reliable solutions to specific problems. Although the approaches appear rather diverse, a detailed analysis reveals that they all share a common basis: the paradigm of the *in silico* experiments (Fig. 1.4). Every time molecular biologists are faced with difficult problems that cannot be addressed by means of standard experimental techniques, novel computational approaches can be designed to provide alternative solutions. *In silico* experiments consist of three stages: (i) methodological development to solve particular problems; (ii) test-case application and comparison with available experimental data; and (iii) *blind* application for real predictions.

Once the biological problem has been carefully analyzed and the limitations of possible experimental methods considered, computational strategies can be proposed. A big advantage of a computational approach is that the inherent model is fully under the control of the user who can optimize it for the specific problem he is dealing with. Hence, before performing the ultimate experiment, methods are usually tested and validated on problems,

whose solution is known. This preliminary phase is very important for the design of a successful protocol: the parameters of the model are fine-tuned, the range of validity of the method estimated and the procedure customized to improve its effectiveness (Fig. 1.4, phase I). When the method is reliable enough, the procedure is applied to a set of more difficult test cases to further challenge the strategy, find its weaknesses and suggest beneficial improvements. Along the whole stage of refinement, the comparison with the experiments is mandatory and represents a *conditio sine qua non* to validate the simulations results (Fig. 1.4, phase II). If any inconsistency is found, the procedure must be reviewed in the light of the new data and further methodological development is required. Through subsequent iterations, *in silico* experiments are continuously refined and improved. At each cycle, the results should become more and more reliable and the description of the biological process more accurate. Upon validation, the global procedure can be applied to the biological problem for which it was originally designed (Fig. 1.4, phase III). This time, computational results become real predictions that can be trusted on the basis of the former testing.

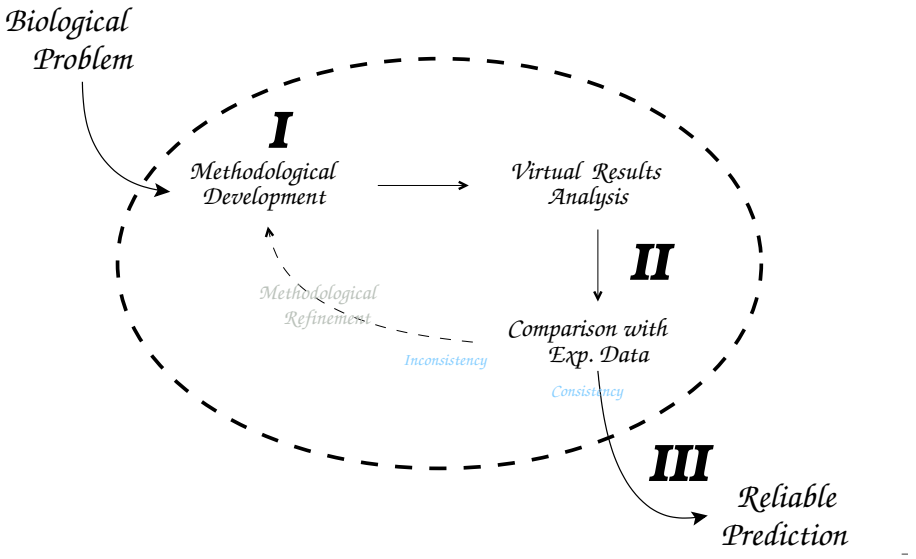


Fig. 1.4: Paradigm of the *in silico* experiments

By following the paradigm shown in Fig. 1.4, two *in silico* experiments have been designed, performed and reported in this thesis. The biological problems behind the experiments, i.e., molecular docking and amyloid peptide aggregation, are intrinsically diverse as are the computational methods developed and used. Nevertheless, both research projects can be included in the common operational framework of the *in silico* experiment.

1.3 Molecular Docking

The identification of molecules that bind the key regions of pharmacologically relevant macromolecules with high affinity and selectivity and alter or inhibit their biological function is known as "drug discovery". Drug discovery is a highly multidisciplinary research process which starts with the identification of a *target* of therapeutic value through biological studies. In order to find putative inhibitors, large molecular libraries are screened and the resulting *leads*, i,e., ligands with high affinity and low molecular weight, are optimized in a cycle that includes design, synthesis, biochemical activity assaying (*in vitro* experiments) and animal studies (*in vivo* experiments).

At the lead optimization stage, the determination of the crystal structure of the target in a complex with some discovered leads is rather important, if not essential. Structural information, which provides the atomic detail of the binding modes of actual ligands, contributes to the understanding of the key elements governing protein-ligand recognition and dramatically speeds up the rational optimization. Unfortunately, the availability of crystal structures for protein-ligand complexes is not always timely and this step may be the "bottleneck" of the entire project. Crystal structure determination is not the only problem that must be faced with during drug discovery. Many hurdles and pitfalls must be overcome before a single molecule that ends up as a drug may be proposed.

At the lead generation stage, experimental high-throughput screening (HTS) requires a molecular library and a reliable activity assay to identify a set of *hits* [38]. A successful *hit*, which is then called a lead compound, must display a potency in the low micromolar range, i.e., the concentration of inhibitor that determines 50% reduction of target activity (IC_{50}) should be 10 μM or less. As soon as a lead compound has been found, extensive optimization and development are required to increase its binding affinity so that the effective concentration gets to the nanomolar range, which is an essential requisite for further development. Sometimes there are no hits at all [39] and a new library has to be selected and screened. As a result of the combination of low *hit* rates and high costs the large-scale approaches of the early 90's are now out of favor [40]. Towards the end of the preclinical period, i.e., when a series of compounds with adequate potency has finally been identified, subtle pharmacological concerns relating to bioavailability, duration of action and toxicity must be addressed. In conclusion, drug discovery is a very difficult and risky process which involves high costs, interdisciplinary expertise and extensive human resources.

Computational Approach

To overcome these limitations, more efficient, faster and cheaper approaches must be pursued. Following the paradigm of the *in silico* experiment (see

previous section), a computer-aided strategy designed to discover new leads from first physico-chemical principles can be proposed. Given the three-dimensional structure of the biomolecular target, the ultimate goal of the procedure is to reproduce the correct binding mode of a small molecule in the binding site of the receptor. the methodology is a valuable help for medicinal chemists who are aiming to discover novel ligands, since it allows them to start the long process of drug development from a sensible point. Computational approaches that mimic the molecular recognition between a target biomolecule and a small compound on a computer have been intensively studied and developed. In the literature, they are known as "molecular docking" techniques [41, 42, 43, 44]. The basic strategy of any docking approach is to generate a conformation of the ligand (*conformer*), to place (or *dock*) it in the active site of the receptor (*binding mode*) and assign a score that will be used to produce a *ranking*. In molecular docking, the essential prerequisites are the availability of a three-dimensional structure of the target, a clever search strategy to sample the conformational space of the ligand and a reliable scoring function to distinguish between active and non-active binding modes. The proper combination of an effective search algorithm and an accurate scoring function are the keys for a successful protocol. When the procedure is applied to a library of available compounds, the approach takes the name of virtual screening (VS) and represents the computational analog of HTS. The library is screened against the target, i.e., each molecule is docked in the receptor's active site, the compounds are ranked according to their binding affinities and a subset of virtual "hits" is suggested for experimental testing. Whereas VS remains less used than HTS for lead discovery, the increased robustness of search algorithms and scoring functions, the availability of affordable computational power, and the potential for timely structural determination of target molecules is making it more practical. Moreover, the speed at which VS can be completed, the low costs, i.e., no need for robotics, reagent acquisition and compound storage facilities, and the potential for screening compounds belonging to available electronic collections make this technique more and more attractive.

*I*n silico Experiment

As for any *in silico* experiment, the paradigm shown in Fig. 1.4 holds true for the VS and three phases must be completed. In the first step (the methodological development), the two components of molecular docking, i.e., the *search strategy* and the *scoring function*, are considered. Docking procedures belong to the category of global optimization techniques. The critical element of the search strategy is the amount of time required to sample the relevant conformational space and find the global minimum of the scoring function, i.e., the search effectiveness. Most optimization algorithms for docking fall into one of three classes: (i) gradient-based al-

gorithms, (ii) combinatorial algorithms and (iii) stochastic algorithms [45]. Gradient-based algorithms are local optimizers and can be used effectively only in combination with other search strategies, such as cycles of Monte Carlo (MC) perturbation and gradient minimizations [46]. Combinatorial algorithms are extremely fast and effective [47, 48, 49, 50, 51] unless the number of conformational degrees of freedom becomes large and the dimension of the search space explodes. Stochastic algorithms have the advantage that irrespective of the dimensionality of the problem, if provided there is sufficient time, they are able to get arbitrarily close to the global minimum. However, the amount of CPU time to achieve an acceptable solution may be relatively large. Although computationally expensive, given the dimensionality of the search space and the ruggedness of the binding energy landscape stochastic algorithms have been shown to be the most suitable optimizers for flexible docking [52, 53, 42, 54, 55, 44]. Several search algorithms have been proposed for docking. Up until now, none of them claims to work in all cases and more effective algorithms are eagerly expected. A brief overview of published strategies is given in **Chapter 2**.

In any docking approach, the scoring function represents the model of protein-ligand interactions. The model should be accurate enough to effectively distinguish the "active" binding mode from all of the others that have been explored and simple enough to permit the evaluation of a large number of potential solutions. Sophisticated energy functions, which accurately describe the thermodynamics of protein-ligand interactions, are computationally too expensive and therefore cannot be used for docking. Furthermore, the resulting energy landscape must be smooth to allow the search to proceed efficiently without becoming trapped in local minima. As suggested by Verkhivker [53], an "adequate" scoring function should fulfill both a thermodynamic and a kinetic requirement. The energy related to the crystallographic structure of the ligand in the complex should be the global minimum of the binding energy landscape (*thermodynamic requirement*). At the same time, this conformation should be accessible during the search (*kinetic condition*). The complexity of a complete and accurate force field that describes the binding process precisely, thus fulfilling the thermodynamic requirement, typically results in a highly frustrated energy landscape and therefore does not meet the kinetic criterion of the docking problem. Hence, simpler molecular recognition models which fulfill both of the requirements need to be designed and developed. Molecular recognition models should at least include both steric and electrostatic terms that account for surface complementarity and hydrogen-bond formation. An intraligand energy term is also required. The latter largely reduces the conformational space to be sampled by preventing intraligand steric clashes. In general, the key elements of a scoring function for robust structural assessment in docking are: (i) protein-ligand steric interaction terms; (ii) a simple description of protein-ligand hydrogen-bonding and (iii) an intramolecular self-avoidance

term. Other contributions, such as dihedral energy and solvation effects, might also play a crucial role in specific cases. Hence, provided that both thermodynamic and kinetic requirements are satisfied, the scoring function should be developed and adapted on a case by case basis, e.g., by considering the physico-chemical properties of the binding site of the target or the nature of the library to be screened. Widespread scoring functions for docking and rationale guidelines for their development are discussed in **Chapter 2**.

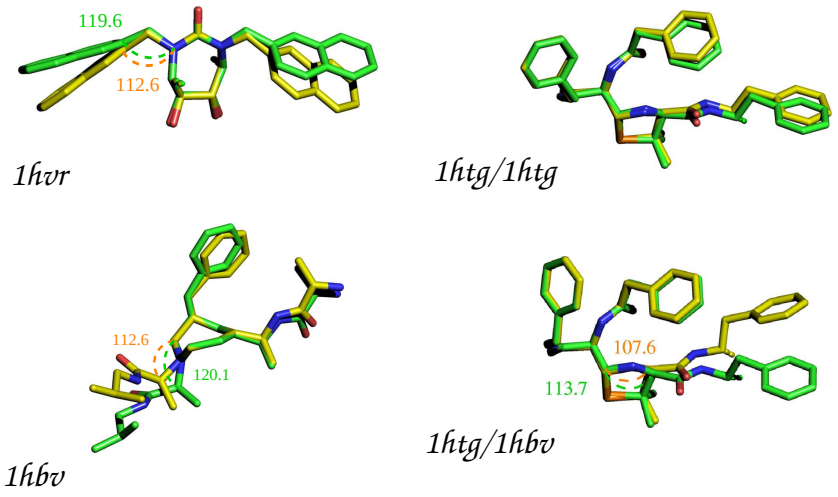


Fig. 1.5: Comparison between ligand structures with biased geometries (green carbons) and unbiased geometries (yellow carbons) used as input for re-docking 1hvr, 1hbv, 1htg and for cross-docking 1htg with the protease 1hbv. Significant deviations in the covalent angles are marked by dashed arcs. (The pictures of the ligands were drawn using the program PyMOL [56]).

In the second stage of the *in silico* experiment, the docking strategy is tested on known potent inhibitors with the aim of reproducing their experimental binding mode. Despite the lack of predictive relevance, redocking experiments represent a very useful tool to challenge the strategy, find weaknesses and suggest beneficial improvements. A "training set" of protein-ligand structures representing the target in complex with different inhibitors is then considered. All ligands are removed from the receptors and minimized when fully solvated, thus removing any bias originating from the crystal structure of the complex. Redocking simulations are carried out and the predicted solutions are compared with the reference, i.e., the binding mode in the crystal structure. To evaluate the reliability of the results, the heavy atom root mean square deviation (RMSD) between the predicted binding modes and the reference are computed. If the redocking experiments are suc-

cessful, i.e., the great majority of low-energy conformations of the ligand are docked within 2.0 Å RMSD from the reference, the strategy is further tested with more difficult training sets, e.g., with more flexible ligands, or more stringent cases, e.g., cross-docking experiments. In the event that the latter tests are successfully passed, the docking strategy is validated and the third phase of the *in silico* experiment is entered. In contrast, if docking predictions do not match the experimental solutions a new stage of methodological development is required. Both methodological refinement and test case applications for the SEED-FFLD strategy [50, 51, 44, 57], the fragment-based docking procedure developed in house, are presented in **Chapter 3**. The SEED-FFLD strategy has been tested on highly flexible inhibitors of human immunodeficiency virus type 1 protease (HIV-1 PR), human α-thrombin and the estrogen receptor β. The docking results indicate that it is possible to correctly reproduce the binding mode of inhibitors with more than ten rotatable bonds, except in cases when the strain in the covalent geometry of the ligand upon binding is large (Fig. 1.5). Hence, automatic approaches that sample only in dihedral space can give misdocked predictions. For docking a limited set of compounds, approaches that allow for full flexibility (including bond angles and lengths) of the ligands, albeit computationally very expensive, should be preferred. Finally, a "hybrid search strategy" consisting of local search and genetic algorithm significantly improves the quality of the SEED-FFLD docking predictions at a moderate additional computational cost (Fig. 1.6). The results of the docking study suggest that hybrid search methods, i.e., combinations of global optimization procedures and local minimization algorithms, should be preferred to global optimization algorithms alone. Thanks to the methodological improvements described in **Chapter 3**, i.e., the introduction of a more efficient and effective search procedure and replacement of the step functions in the protein-ligand polar term with continuous bathtub-shaped profiles, the in-house fragment-based docking strategy could be used in computational screening projects to discover novel inhibitors against relevant targets. Upon ranking, the top scoring compounds have been purchased or synthesized and submitted for experimental testing. Numerous successes of this approach are well documented in the literature [58, 59]. In our laboratory, the SEED-FFLD strategy has been successfully applied in a VS project against β-secretase, a very difficult target involved in Alzheimer's disease [60]. By screening a library of about 300000 molecules (the iResearch library) and a library tailored for proteases, thirty compounds were predicted to have affinities in the high nanomolar range. Most of them were phenylurea derivatives. Upon refinement, the good candidates were purchased and tested *in vitro* and by two cell-based assays. Interestingly, three compounds showed low-micromolar activity in the cell based-assays. Given the the very small size (MW=322) of one of the three, this compound is a very promising lead candidate for β-secretase inhibition. In a second screening experiment against the same target eighty-

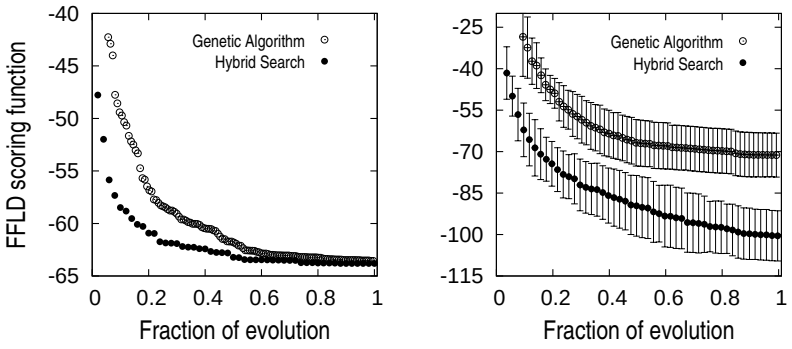


Fig. 1.6: Evolution of the best individual of the population averaged over ten docking runs for two different experiments. Empty and filled bullets indicate evolutions performed by genetic algorithm and hybrid search procedure, respectively. Docking of HIV-1 proteinase ligands with 10 and 21 rotatable bonds are shown in the left- and right-hand plots, respectively. In the right-hand plot, the vertical bars show the standard deviation computed over ten docking runs.

eight compounds suggested *in silico* were tested *in vitro*, and 10 of them showed an IC50 value lower than 100 μM in a BACE-1 enzymatic assay. The 10 active compounds shared a triazine scaffold. Moreover, four of them were active in an assay with mammalian cells (EC50 ¡ 20 μM), indicating that they are cell-permeable. Such as the phenylurea compounds, these triazine derivatives are very promising lead candidates for BACE-1 inhibition. The details of the two VS experiments, the *in vitro* validation of the computational predictions and the identified active compounds are described in **Chapter 4** and **Chapter 5**.

1.4 Amyloid Aggregation

The term "amyloid" was originally introduced to describe certain deposits found *post-mortem* in organs and tissues, which gave a positive reaction when stained with iodine [61]. It was realized only later that the amyloid material was predominantly proteinaceous. With the increasing precision in the definition of amyloid due to its characteristic green birifrengence when stained with the dye Congo Red [62], its particular appearance in the electron microscope [63] and its specific X-ray diffraction pattern [64], it has become evident that amyloid deposits are highly-ordered aggregates in which the polypeptide units lie in a "non-native" conformation. Thus, amyloid aggregation is tightly linked to protein misfolding, i.e., the inability of a protein to fold correctly or remain correctly folded so that under cer-

tain conditions it assumes a different three-dimensional structure that can induce the fibril formation.

The major interest in amyloids arises from the evidence that ordered deposits are associated with a number of severe human disorders (*amyloidoses*) including Alzheimer's disease, Hungtington's disease, transmissible spongiform encephalopathies, i.e., bovine spongiform encephalopathy (BSE), Creutzfeldt-Jakob disease (CJD) and Kuru, type II diabetes and a number of systemic polyneuropathies [65]. Presently, more than 20 proteins are known to be responsible for diverse human amyloidoses. The polypeptide chains involved in amyloid diseases include full-length proteins (e.g. lysozyme or immunoglobulin light chains), biological peptides (amylin, atrial natriuretic factor) and fragments of larger proteins produced either by specific processing (e.g. the amyloid-β peptide) or by more general degradation (e.g. poly-Q stretches cleaved from proteins with poly-Q extensions such as hungtintin, ataxin and the androgen receptor). In some cases amyloid deposits involve wild-type sequences and, in other cases, variants resulting from genetic mutations. This latter group is associated with familial forms of the disease which are particularly aggressive and usually correspond to earlier onsets [66]. The soluble precursors of the ordered deposits do not share any sequence homology or common fold.

Despite the apparent diversity, the presence in tissue of proteinaceous deposits is a hallmark of amyloidoses, thus suggesting a causative link between aggregate formation and pathological symptoms (the amyloid hypothesis) [67, 68, 69]. Moreover, X-ray diffraction data indicate a common cross-β structure for most fibrillar aggregates [70, 71]. The latter finding suggests, on one hand, that the key steps in aggregation may be common to all amyloid proteins and, on the other hand, that fibrillar or pre-fibrillar structures may be the origin of the disease. Hence, to understand the molecular basis of the diseases, the structure of the aggregates should be carefully analyzed. In the last 50 years, a lot of effort has been put into the structural determination of amyloid aggregates. Initially, these deposits have been characterized by staining methods, electron microscopy and X-ray diffraction. The early investigations indicated that amyloids, regardless of the disease, share the following structural features:

- under electron microscope, amyloid deposits can be seen to be composed of uniform, straight, unbranched fibers, $\sim$ 100 Å in diameter and of indefinite length [72]; the fibers are usually straight or only slightly curved suggesting that they have a particularly rigid molecular structure (Fig. 1.7);

- the molecular structure is such that amyloid fibrils bind Congo Red dye and interact with this planar *bis*-diazo dye in such a way that the bound molecules are spatially ordered with respect to the fibril and generate a characteristic and diagnostic green birefringence [73];

- amyloid X-ray diffraction patterns show that the repeating molecular structure of the fibrils consists of polypeptide chains in extended β-conformation, hydrogen-bonded together into sheets which run perpendicular to the axis of the fibril, the so called "cross-β" conformation [74, 75].

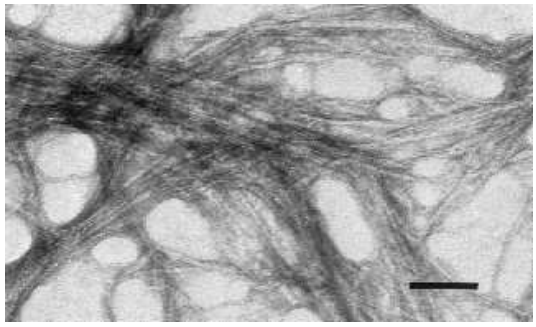


Fig. 1.7: Electron micrograph of amyloid-β fibrils

The "cross-β" structure unveiled by X-ray studies is rather intriguing because it incorporates an inherent propensity for propagation. In cross-β conformation, polypeptide chains are extended with alternate peptide groups rotated by 180°. Each carbonyl and amide group in any β-strand can make hydrogen bonds with the amide or carbonyl group, respectively, on parallel or antiparallel adjacent strands. In turn, adjacent strands can make hydrogen bonds with other similarly organized strands and, in principle, an indefinite number of monomeric units can be included. To understand the mechanism of amyloid formation and the nature of the energetic contributions that stabilize similar ordered arrangements for such a diverse class of proteins, atomic resolution structures would be required. However, no high-resolution three-dimensional structure of an amyloid fibril has been determined yet. Amyloid fibrils are noncristalline solid materials and therefore highly incompatible with the usual techniques for protein structure determination, i.e., X-ray crystallography and liquid state NMR.

To obtain structural information at molecular level, more sophisticated approaches have to be followed. Recent experiments by Lansbury, Griffin and co-workers [76, 77] and by Linn, Meredith, Botto and co-workers [78, 79] have demonstrated that structural constraints on amyloid fibrils can be obtained from solid-state NMR, which requires neither crystallinity nor solubility. By appropriate isotopic labelling (e.g., ^{13}C or ^{15}N), solid state NMR measurements provide quantitative data on the supramolecular organization of the β-sheets in the fibril and describe protein fibrillar conformations in terms of torsion angles or interatomic distances. Remarkably, measurements of ^{13}C–^{13}C nuclear magnetic dipole-dipole couplings have provided

experimental evidence of the in-register parallel alignment of the amyloid-β peptide in the deposits associated with the Alzheimer's disease [80, 81, 82]. Although these technique are useful, they can supply "only" a certain number of constraints. Experimental results may be insufficient to define a unique protein fibrillar conformation and the resulting structural models may be rather inaccurate, if not incorrect. More importantly, the specific nature of the pathogens in amyloidoses and the basis of cytotoxicity are still unclear and the subject of an intense debate [83, 84, 85, 86, 87, 88, 89]. An increasing quantity of recent experimental data suggest that the early pre-fibrillar aggregates are the most toxic species [90, 91]. On the hypothesis that prominent cytotoxicity is exhibited by the soluble precursors of amyloid fibrils, structural information on mature fibrils cannot be used to prevent amyloid diseases. Hence, alternative approaches providing structural and dynamic information should also be pursued in addition to solid state NMR experiments.

Computational Approach

To study amyloid aggregation at atomic detail, alternative approaches involving computational methods may be invoked. In this case, the ultimate goal of the *in silico* experiment would be to monitor the self-assembly of a certain number of amyloid proteins and determine the complete free-energy surface of the oligomeric system. This would also provide structural characterization of the early aggregates, which are believed to be cytotoxic. The nature of the energetic contributions that stabilize ordered aggregates could be identified and the molecular basis of fibril formation explained. Moreover, by testing the *in silico* behavior of engineered mutants the sequence dependence of amyloidogenicity, i.e., the capability of forming amyloid fibrils, could be easily investigated.

However, *in vitro* fibril formation is rather slow and the time taken for this process ranges from several minutes to days, dependent on which amyloid sequence we are dealing with. Such timescales are never accessible by standard computer simulations. Moreover, the process is intrinsically cooperative and several aggregating units must be included in the model. For these reasons, the size of the system, i.e., the total number of interacting centers, becomes intractable very rapidly. To keep the complexity and CPU requirements low, only "small" and "slightly realistic" oligomeric systems can be investigated. Recent all-atom MD simulations of a three amyloidogenic heptapeptides `GNNQQNY` have shown that the effective energy surface of such a simplistic aggregating system is already highly rugged [92]. The absence of connections between the peptides results in a large number of different low-energy states, which increases the frustration of the system and hinders the sampling. Since current simulation approaches only allow significant sampling for small oligomers (e.g., typically those with no more

than six peptide replicas with fewer than ten residues each) larger and "more relevant" systems cannot be investigated. These limitations strongly reduce the effectiveness of the computational approach. Indeed, many important questions about the formation of ordered aggregates have been addressed by computational studies. Simplified on-lattice models have allowed to investigate the foldability and aggregation propensity [93, 94] and how interaction potentials affect the properties of aggregation-prone proteins [95]. Other studies have shown the less stable the protein, the greater the chance that it will assume an alternative native state as multimer [96]. A minimalist "Go" model of four peptide strands [97] has been investigated by MD simulations in a confining sphere and the aggregation process was shown to be dependent on both sequence and environment [98].

*I*n silico Experiment

Following the paradigm shown in Fig. 1.4, the optimal computational protocol is designed to reduce the *statistical errors* originating from the timescale of the process and the frustration of the system as much as possible (phase I). Hence, "simple" amyloid systems can be investigated by applying clever sampling procedures, such as the replica exchange molecular dynamics (REMD) approach [99, 100], and simplified energy models, such as implicit treatments of the solvent [101, 102, 103, 35]. The former is an efficient way to simulate complex systems at physiological temperatures, the latter is a continuum representation of the solvent that drastically reduces the computational cost of the simulations and allows the investigations to have much longer timescales (in the μs range).

"Replica exchange" is the simplest form of simulated tempering [99]. Sugita and Okamoto were the first to extend the original formulation of replica exchange into an MD-based version and they tested it on the pentapeptide Met-enkephalin *in vacuo* [100]. Several atomistic simulations of proteins have shown the efficiency of the method [104, 105, 106, 107, 108, 109, 110]. The basic idea of REMD is to simulate different copies (*replicas*) of the system at the same time but at different temperatures values. Each replica evolves independently by MD and, every $t_textrmswap$ time interval, states i, j which have neighboring temperatures are swapped (by velocity rescaling) with a probability $w_{ij} = \exp(-\Delta)$, [100] where $\Delta \equiv (\beta_i - \beta_j)(E_j - E_i)$, $\beta = 1/kT$ and E is the effective energy (potential and solvation energy). The result of this swapping between different temperatures is that high temperature replicas help the low temperature ones to jump across the energy barriers of the system. During the simulation, each replica visits all the temperatures of the set several times, thus undergoing a free random walk in temperature space [100]. This, in turn, corresponds to a random walk in energy space that enhances the sampling. Simple protocols to set up a REMD simulation of amyloid peptide aggregation are described in **Chapter 6**.

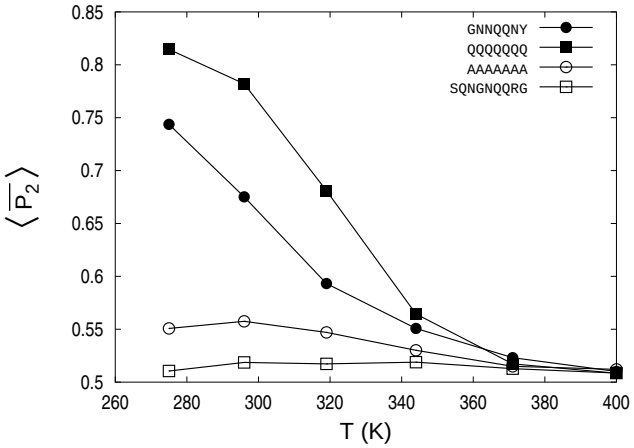


Fig. 1.8: Temperature dependence of the nematic order parameter $\langle \overline{P_2} \rangle$ averaged over the canonical ensembles sampled by REMD for four oligomeric peptide systems. $\langle \overline{P_2} \rangle$ estimates the amyloidogenic propensity of peptide systems and discriminates between amyloidogenic (GNNQQNY and QQQQQQQ) and non amyloidogenic (SQNGNQQRG and AAAAAAA) sequences in agreement with experimental data [112, 113, 114].

Our study showed that REMD samples conformation space and aggregation transitions more efficiently than constant temperature molecular dynamics (CTMD) at physiologically relevant temperature values [111]. To simulate amyloid peptide aggregation, the following protocol can be adopted: each simulation run is carried out with three peptide replicas starting from random conformations, positions, and orientations. In the initial random positions, there must not be any intermolecular contact, i.e., the peptides are separated in space. The system is simulated in a cubic box, whose sides are adjusted to yield the desired sample concentration ($\sim$ 5 mg/ml to be consistent with *in vitro* experiments), and with periodic boundary conditions. As soon as some thermodynamic observable which has been selected to monitor aggregation (e.g., such as the fraction of parallel or antiparallel interstrand contacts) has reached convergence the simulation is stopped. For the systems we studied (three 7-residue peptide replicas simulated by REMD), this corresponded to a 1μs simulation time which requires approximately 2 weeks on a current processor. Once the simulations are finished, the trajectories are analyzed and the results compared with available experimental data (Fig. 1.4, phase II). To describe the aggregation process in terms of free-energy profiles and surfaces, appropriate progress variables must be chosen. In the simulation study reported in **Chapter 6**, two progress variables were defined: the radius of gyration of the oligomeric system R_g and the nematic order parameter $\overline{P_2}$: the former monitors the degree of *condensation* of the system, the latter the aggregation process viewed as an order transition.

The computational results showed that it is possible to simulate with an atomic model the early steps of aggregation of amyloid forming peptides and obtain high-resolution structural information (Fig. 1.10). Moreover, the early steps have been interpreted as a condensation step leading to the formation of disordered aggregates followed by an order transition, in agreement with experimental evidence [115]. Finally, the nematic $\overline{P_2}$ averaged over the canonical ensemble, which is referred to as β-aggregation propensity, can effectively estimate the amyloidogenic propensity of the system and discriminate amyloidogenic from soluble peptides in agreement with experimental data [113, 112, 114] (Fig. 1.8).

Given the substantial agreement between simulation predictions and experimental measurements, the phase III of the *in silico* experiment has been entered to investigate the aggregation properties of the Alzheimer's human amyloid-β peptide ($A\beta_{42}$). Due to its large size (42 residues), oligomeric systems of full length amyloid-β peptides cannot be effectively studied by computer simulations. Hence, a novel method to investigate the aggregating properties of amyloid proteins has been developed. The strategy consists of four steps: (i) the amyloid protein is dissected into sets of short overlapping stretches, i.e., 7- or 11-residue long, that encompass the full-length sequence; (ii) for every stretch, the aggregation process of small oligomeric systems, i.e., between three and six peptide replicas, is simulated by all-atom MD; (iii) the β-aggregation propensity is computed and (iv) used to build the amyloidogenicity profile, i.e., the position dependence of β-aggregation propensity along the protein sequence. In summary, the amyloidogenicity profile is obtained by dissecting the whole amyloid protein into short overlapping stretches and computing the aggregation propensity of each segment. Based on the hypothesis of in-register parallel arrangements, aggregation MD simulations should reproduce the fibrillar environment experienced by each segment. The observations made on the stretches can then be extended to the full-length protein and its aggregating properties deduced with reason. The MD-based amyloidogenicity profile on $A\beta_{42}$ highlighted the region from Val12 to Asp22 as the major aggregation *hot-spot*. Although with a lower tendency, the C-terminal segment (residues 31-37) was also found to be aggregation-prone. These findings are in good agreement with radioligand binding experiments [116] and ThT-fluorescence assays [117] and are rather consistent with the structural model for $A\beta_{40}$ fibrils derived from solid state NMR [118]. The enhanced β-aggregation propensity detected at the N-terminus by the implicit solvent runs is rather surprising and in disagreement with solid state NMR [118] and EPR measurements [119]. It is likely that the approximations inherent to the solvation model, and in particular the neutralization of formal charges, were too crude to correctly reproduce the behavior of polypeptide segments with many charged side chains. Explicit solvent simulations started from parallel β-sheet conformations of segments located at the N-terminus unveiled their marginal

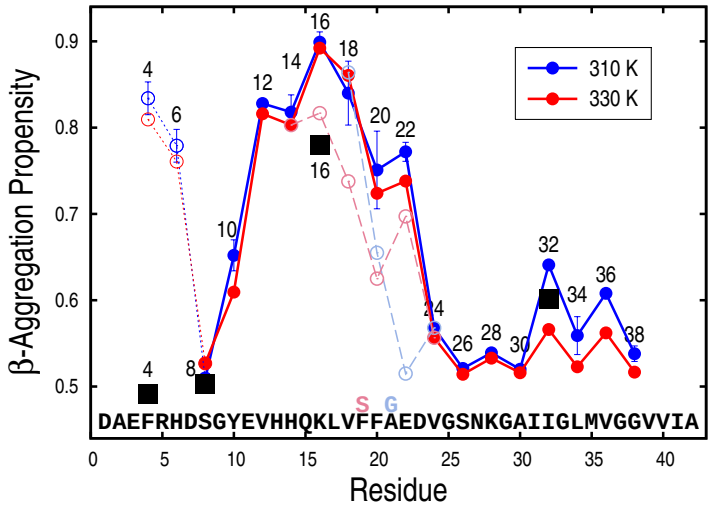


Fig. 1.9: Results of constant temperature MD simulations of trimeric 7-residue peptide systems. Values of the β-aggregation propensity along the $A\beta_{42}$ sequence at 310 (blue) and 330 K (red) obtained from aggregation simulations of three 7-residue peptides.

structural stability (see black squares in Fig 1.9), in agreement with experiments. Thanks to the atomic detail provided by the MD simulations, the β-aggregation profile could be interpreted on a structural basis. A secondary structural analysis unveiled the presence of four potential β-turn or bend sites along the amyloid-β sequence: "SG" (res. 8-9), "GS" (res. 25-26), "GA" (res. 29-30), "GV" (res. 38-39). The four potential β-turns correspond to sudden drops in the β-aggregation profile and are located at the borders of the identified aggregation-prone regions. The simulation results suggest that these four potential β-turns play a critical role in determining the aggregation properties, i.e., the location of the aggregation *hot-spots*, the amyloidogenic content, i.e., the capability of forming amyloid fibrils, and the fibrillar structure of the amyloid-β peptide. A detailed discussion of these simulation results is reported in **Chapter 7**. Experimental validation of the *in silico* prediction is mandatory and represents a challenge that has been left to experimentalists.

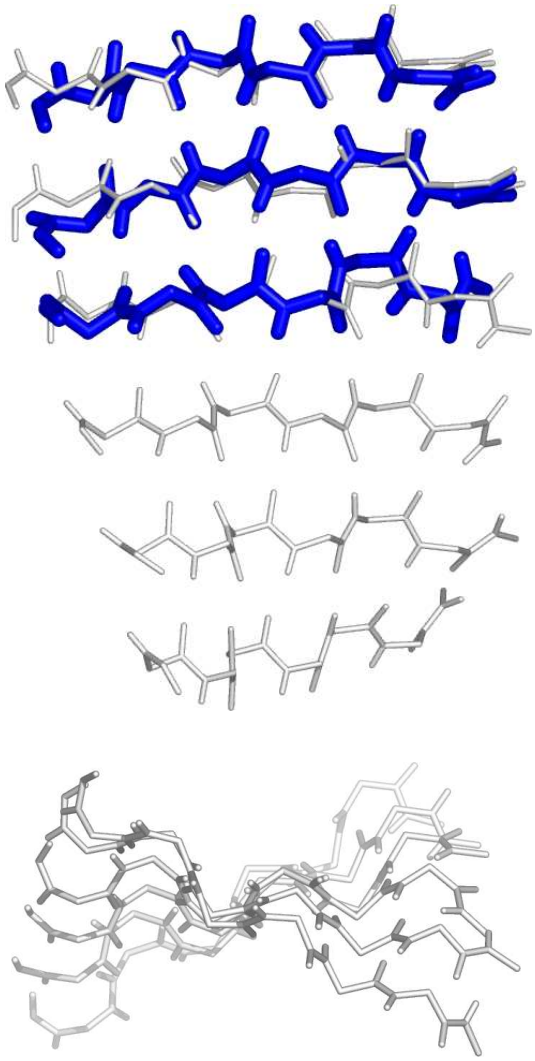


Fig. 1.10: (Top) Snapshots of ordered aggregates of three (thick sticks) and six (thin sticks) amyloidogenic SYVIIE peptides [120] extracted from CTMD simulations at 330 K. The simulations were performed at a sample concentration of 5 mg/ml. The overall conformation and twist of the three-stranded and six-stranded parallel β-sheets are indistinguishable. (Bottom) The six-stranded β-sheet upon 90° rotation to better visualize the twist. (The pictures were drawn using the program PyMOL [56]).

Bibliography

[1] Anfinsen, C. B. (1973) *Science 181*, 223–230.

[2] Lacroix, E., Viguera, A. R. & Serrano, L. (1998) *Folding & Design 3*, 79–85.

[3] Fitch, W. M. & Margoliash, E. (1970) *Evol. Biol. 4*, 67.

[4] Chothia, C. & Lesk, A. M. (1986) *EMBO J. 5*, 823–827.

[5] Dinner, A. R., Sali, A., Smith, L. J., Dobson, C. M. & Karplus, M. (2000) *Trends in Biochemical Sciences 25*, 331–339.

[6] Dill, K. & Chan, H. (1997) *Nature Struct. Biol. 4*, 10–19.

[7] Dobson, C. M., Šali, A. & Karplus, M. (1998) *Angew. Chem. Int. Ed. 37*, 869–893.

[8] Levine, R. E. & Bernstein, R. B. *Molecular Reaction Dynamics and Chemical Reactivity.* Oxford University Press, Oxford, (1987).

[9] Herschbach, D. R. (1987) *Angew. Chem. Int. Ed. 26*, 1221–1243.

[10] Abrahams, J. P., Leslie, A. G. W., Lutter, R. & Walker, J. E. (1994) *Nature 370*, 621–628.

[11] Groll, M., Ditzel, L., Lowe, J., Stock, D., Bochtler, M., Bartunik, H. D. & Hüber, R. (1997) *Nature 386*, 463–471.

[12] Ban, N., Ditzel, L., Lowe, J., Stock, D., Bochtler, M., Bartunik, H. D. & Hüber, R. (2000) *Science 289*, 905–920.

[13] Wimberly, B. T., Brodersen, D. E., Clemons, W. M., Morgan-Warren, R. J., Carter, A. P., Vonrhein, C., Hartsch, T. & Ramakrishnan, V. (2000) *Nature 407*, 327–339.

[14] Vendruscolo, M., Zurdo, J., MacPhee, C. E. & Dobson, C. M. (2003) *Phil. Trans. R. Soc. A 361*, 1205–1222.

[15] Plaxco, K. W. & Dobson, C. M. (1996) *Curr. Opin. Struct. Biol. 6*, 630–636.

[16] Eaton, W. A., Munoz, V., Thompson, P. A., Chan, C. K. & Hofrichter, J. (1997) *Curr. Opin. Struct. Biol. 7*, 10–14.

[17] Williams, S., Causgrove, T. P., Gilmanshin, R., Fang, K. S., Callender, R. H., Woodruff, W. H. & Dyer, R. B. (1996) *Biochemistry 35*, 691–697.

[18] Weiss, S. (1999) *Science 283*, 1676–1683.

[19] Mehta, A. D., Rief, M., Spudich, J. A., Smith, D. A. & Simmons, R. D. (1999) *Science 283*, 1689–1695.

[20] Ishii, Y. & Yanagida, T. (2000) *Single Molecules 1*, 5–13.

[21] Funatsu, T., Harada, Y., Tokunaga, M., Saito, K. & Yanagida, T. (1995) *Nature 374*, 555–559.

[22] Ishijima, A., Kojima, H., Funatsu, T., Tokunaga, M., Higuchi, H., Tanaka, H. & Yanagida, T. (1998) *Cell 92*, 161–171.

[23] Harada, Y., Harada, Y., Funatsu, T., Murakami, K., Nonoyama, Y., Ishihama, A. & T., Y. (1999) *Biophys. J. 76*, 709–715.

[24] Sako, Y., Minoghchi, S. & T., Y. (2000) *Nat. Cell Biol. 2*, 168–172.

[25] Sako, Y., , Minoghchi, S. & T., Y. (2000) *Single Molecules 2*, 151–155.

[26] Fersht, A. *Structure and mechanism in protein science: a guide to enzyme catalysis and protein folding.* W. H. Freeman and Co., New York, (1999).

[27] Vitkup, D., Ringe, D., Petsko, G. A. & Karplus, M. (2000) *Nature Struct. Biol. 7*, 34–38.

[28] Xu, A., Horwich, S. C. & Sigler, P. (1997) *Nature 388*, 741–750.

[29] Ma, J. P., Sigler, P., Xu, Z. H. & Karplus, M. (2000) *J. Mol. Biol. 302*, 303–313.

[30] Bursulaya, B. D. & Brooks III, C. L. (1999) *J. Am. Chem. Soc. 121*, 9947–9951.

[31] Ferrara, P. & Caflisch, A. (2000) *Proc. Natl. Acad. Sci. USA. 97*, 10780–10785.

[32] Snow, Y. M., Nguyen, N., Pande, V. & Gruebele, M. (2002) *Nature 42*, 102–106.

[33] Jang, S., Shin, S. & Pak, Y. (2002) *J. Am. Chem. Soc. 124*, 4976–4979.

[34] Rao, F., Settanni, G., Guarnera, E. & Caflisch, A. (2005) *J. Chem. Phys. 122*, 184901.

[35] Ferrara, P., Apostolakis, J. & Caflisch, A. (2002) *Proteins: Structure, Function and Genetics 46*, 24–33.

[36] Vendruscolo, M., Paci, E., Dobson, C. M. & Karplus, M. (2001) *Nature 409*, 641–645.

[37] Paci, E., Vendruscolo, M., Dobson, C. M. & Karplus, M. (2002) *J. Mol. Biol. 324*, 151–163.

[38] Hertzberg, R. P. & Pope, A. J. (2000) *Curr. Opin. Chem. Biol. 4*, 445.

[39] Böhm, H. J. *et al.* (2000) *J. Med. Chem. 43*, 2664–2674.

[40] Lahana, R. (1999) *Drug Discov. Today 4*, 447–448.

[41] Kuntz, I., Blaney, J., Oatley, S., Langridge, R. & Ferrin, T. (1982) *J. Mol. Biol. 161*, 269–288.

[42] Morris, G., Goodsell, D., Halliday, R., Huey, R., Hart, W., Belew, R. & Olson, A. (1998) *J. Comput. Chem. 19*(14), 1639–1662.

[43] Claussen, H., Buning, C., Rarey, M. & Lengauer, T. (2001) *J. Mol. Biol. 308*, 377–395.

[44] Budin, N., Majeux, N. & Caflisch, A. (2001) *Biol. Chem. 382*, 1365–1372.

[45] Diller, J. D. & Verlinde, C. L. M. J. (1999) *J. Comput. Chem. 20*(16), 1740–1751.

[46] Caflisch, A., Niederer, P. & Anliker, M. (1992) *Proteins: Structure, Function and Genetics 13*, 223–230.

[47] Miller, M., Kearsley, S. K., Underwood, D. J. & Sheridan, M. D. (1994) *J. Comput.-Aided Mol. Design 8*, 153.

[48] Rarey, M. & Stahl, M. (2001) *J. Med. Chem. 44*, 1035–1042.

[49] Makino, S. & Kuntz, I. D. (1997) *J. Comput. Chem. 18*, 1812.

[50] Majeux, N., Scarsi, M., Apostolakis, J., Ehrhardt, C. & Caflisch, A. (1999) *Proteins: Structure, Function and Genetics 37*, 88–105.

[51] Majeux, N., Scarsi, M. & Caflisch, A. (2001) *Proteins: Structure, Function and Genetics 42*, 256–268.

[52] Gehlhaar, D. K., Verkhivker, G. M., Rejto, P. A., Sherman, C. J., Fogel, D. B., Fogel, L. J. & Freer, S. T. (1995) *Chem. Biol. 2*, 317–324.

[53] Verkhivker, G. M., Rejto, P. A., Gelhaar, D. K. & Freer, S. T. (1996) *Proteins: Structure, Function and Genetics 25*, 342–353.

[54] Goldberg, D. E. *Genetic Algorithms in Search Optimization and Machine Learning*. Addison-Wesley, Reading MA, (1989).

[55] Davis, L. *Handbook of Genetic Algorithms*. Van Nostrand Reinhold, New York NY, (1991).

[56] DeLano, W. *The PyMOL Molecular Graphics System*. DeLano Scientific, San Carlos, CA, USA, (2002).

[57] Cecchini, M., Kolb, P., Majeux, N. & Caflisch, A. (2004) *J. Comput. Chem. 25*, 412–422.

[58] Shoichet, B. K., McGovern, S. L., Wei, B. & Irwin, J. J. (2002) *Curr. Opin. Chem. Biol. 6*, 439.

[59] Blake, J. F. & Laird, E. R. (2003) *Ann. Rep. Med. Chem. 38*, 305.

[60] Huang, D., Lüthi, U., Kolb, P., Edler, K., Cecchini, M., Audetat, S., Barberis, A. & Caflisch, A. (2005) *J. Med. Chem.* . in press.

[61] Virchow, R. (1854) *Virchows Arch. 6*, 415–426.

[62] Missmahl, H. P. & Hartwig, M. (1953) *Virchows Arch. Path. Anat. 324*, 489–508.

[63] Cohen, A. S. & Calkins, E. (1959) *Nature 183*, 1202–1203.

[64] Eanes, E. D. & Glenner, G. G. (1968) *J. Histochem. Cytochem. 16*, 673–677.

[65] Pepys, M. B. *et al.* (1994) *Nature 362*, 553–557.

[66] Nilsberth, C. *et al.* (2001) *Nat. Neurosci. 4*, 887–893.

[67] Kelly, J. (1998) *Curr. Opin. Struct. Biol. 8*, 101–106.

[68] Dobson, C. M. (2001) *Phil. Trans. R. Soc. Lond. B 356*, 133–145.

[69] Reilly, M. M. (1998) *J. Neurol. 245*, 6–13.

[70] Blake, C. & Serpell, L. (1996) *Structure (London) 4*, 989–998.

[71] Malinchik, S. B., Inouye, H., Szumowski, K. E. & Kirschner, D. A. (1998) *Biophys. J. 74*, 537–545.

[72] Cohen, A., Shirahama, T. & Skinner, M. Electron microscopy of amyloid. in *Electron microscopy of Protein,* (Harris, I., ed), 165–205. Academic Press, London, (1981).

[73] Glenner, G. & Eanes, E. (1972) *J. Histochem. Cytochem. 20*, 821–826.

[74] Glenner, G. (1980) *New England J. Med. 302*, 1283–1292.

[75] Glenner, G. G. (1980) *New England J. Med. 302*, 1333–1343.

[76] Griffiths, J., Ashburn, T., Auger, M., Costa, P., Griffin, R. & Lansbury, P. (1995) *J. Am. Chem. Soc. 117*, 3539–3546.

[77] Jaroniec, C., MacPhee, C., Bajaj, V., McMahon, M., Dobson, C. & R.G., G. (2004) *Proc. Natl. Acad. Sci. USA. 101*, 711–716.

[78] Burkoth, T., Benzinger, T., Urban, V., Morgan, D., Gregory, D., Thiyagarajan, P., Botto, R., Meredith, S. & Lynn, D. (2000) *J. Am. Chem. Soc. 122*, 7883–7889.

[79] Benzinger, T., Gregory, D., Burkoth, T., Miller-Auer, H., Lynn, D., Botto, R. & Meredith, S. (2000) *Biochemistry 39*, 3491–3499.

[80] Antzutkin, O. N., Balbach, J. J., Leapman, R. D., Rizzo, N. W., Reed, J. & Tycko, R. (2000) *Proc. Natl. Acad. Sci. USA. 97*, 13045–13050.

[81] Antzutkin, O. N., Leapman, R. D., Balbach, J. J. & Tycko, R. (2002) *Biochemistry 41*, 15436–15450.

[82] Balbach, J. J., Petkova, A. T., Oyler, N. A., Antzutkin, O. N., Gordon, D. G., Meredith, S. C. & Tycko, R. (2002) *Biophys. J. 83*, 1205–1216.

[83] Lambert, M. *et al.* (1998) *Proc. Natl. Acad. Sci. USA. 95*, 6448–6453.

[84] Hartley, D., Walsh, D., Ye, C., Diehl, T., Vasquez, S., Vassilev, P., Teplow, D. & Selkoe, D. (1999) *J. Neurosci. 19*, 8876–8884.

[85] Walsh, D., Hartley, D. M., Kusumoto, Y., Fezoui, Y., Condron, M., Lomakin, A., Benedek, G., Selkoe, D. & Teplow, D. (1999) *J. Biol. Chem. 274*, 25945–25952.

[86] Monji, A., Yoshida, I., Tashiro, K., Hayashi, Y., Matsuda, K. & Tashiro, N. (2000) *Neurosci. Lett. 278*, 81–84.

[87] Goldberg, M. & Lansbury, P. (2000) *Nat. Cell Biol. 2*, E115–E119.

[88] Conway, K., Lee, S., Rochet, J., Ding, T., Williamson, R. & Lansbury, P. (2000) *Proc. Natl. Acad. Sci. USA. 97*, 571–576.

[89] Walsh, D., Klyubin, I., Fadeeva, J., Cullen, W., Anwyl, R., Wolfe, M., Rowan, M. & Selkoe, D. (2002) *Nature 416*, 535–539.

[90] Bucciantini, M. *et al.* (2002) *Nature 416*, 507–511.

[91] Cleary, J., Walsh, D., Hofmeister, J., Shankar, G., Kuskowski, M., Selkoe, D. & Ashe, K. (2005) *Nat. Neurosci. 8*, 79–84.

[92] Gsponer, J., Habertür, U. & Caflisch, A. (2003) *Proc. Natl. Acad. Sci. USA. 100*, 5154–5159.

[93] Broglia, R. A., Tiana, G., Pasquali, S., Roman, H. E. & Vigezzi, E. (1998) *Proc. Natl. Acad. Sci. USA. 95*, 12930–12933.

[94] Bratko, D. & Blanch, H. W. (2003) *J. Chem. Phys. 118*, 5185–5194.

[95] Giugliarelli, G., Micheletti, C., Banavar, J. R. & Maritan, A. (2000) *J. Chem. Phys. 113*, 5072–5077.

[96] Harrison, P. M., Chan, H. S., Prusiner, S. B. & Cohen, F. E. (1999) *J. Mol. Biol. 286*, 593–606.

[97] Vekhter, B. & Berry, R. S. (1999) *J. Chem. Phys. 110*, 2195–2201.

[98] Friedel, M. & Shea, J. E. (2004) *J. Chem. Phys. 120*, 5809–5823.

[99] Marinari, E. & Parisi, G. (1992) *Europhys. Lett. 19*, 451.

[100] Sugita, Y. & Okamoto, Y. (1999) *Chemical Physics Letters 314*, 141–151.

[101] Still, W. C., Tempczyk, A., Hawley, R. C. & Hendrickson, T. (1990) *J. Am. Chem. Soc. 112*, 6127–6129.

[102] Roux, B. & Simonson, T. (1999) *Biophysical Chemistry 78*, 1–20.

[103] Lazaridis, T. & Karplus, M. (1999) *Proteins: Structure, Function and Genetics 35*, 133–152.

[104] Sanbonmatsu, K. & Garcia, A. (2002) *Proteins: Structure, Function and Genetics 46*, 225–234.

[105] Garcia, A. E. & Sanbonmatsu, K. (2002) *Proc. Natl. Acad. Sci. USA. 99*, 2782–2787.

[106] Garcia, A. E. & Sanbonmatsu, K. (2001) *Proteins: Structure, Function and Genetics 42*, 345–354.

[107] Zhou, R., Berne, B. & Germain, R. (2001) *Proc. Natl. Acad. Sci. USA. 98*, 14931–14936.

[108] Pitera, J. W. & Swope, W. (2003) *Proc. Natl. Acad. Sci. USA. 100*, 7587–7592.

[109] Rao, F. & Caflisch, A. (2003) *J. Chem. Phys. 119*, 4035–4042.

[110] Im, W. & Brooks III, C. L. (2004) *J. Mol. Biol. 337*, 513–519.

[111] Cecchini, M., Rao, F., Seeber, M. & Caflisch, A. (2004) *J. Chem. Phys. 121*(21), 10748–10756.

[112] Balbirnie, M., Grothe, R. & Eisenberg, D. (2001) *Proc. Natl. Acad. Sci. USA. 98*, 2375–2380.

[113] Perutz, M. F., Johnson, T., Suzuky, M. & Finch, J. T. (1994) *Proc. Natl. Acad. Sci. USA. 91*, 5355–5358.

[114] Perutz, M. F., Pope, B. J., Owen, D., Wanker, E. E. & Scherzinger, E. (2002) *Proc. Natl. Acad. Sci. USA. 99*, 5596–5600.

[115] Serio, T. R., Cashikar, A. G., Kowal, A. S., Sawicki, G. J., Moslehi, J. J., Serpell, L., Arnsdorf, M. F. & Lindquist, S. L. (2000) *Science 289*, 1317–1321.

[116] Tjernberg, L. O., Näslund, J., Lindqvist, F., Johansson, J., Karlstrom, A. R., Thyberg, J., Terenius, L. & Nordstedt, C. (1996) *J. Biol. Chem. 271*, 8545–8548.

[117] Liu, R., McAllister, C., Lyubchenko, Y. & Sierks, M. R. (2003) *J. Neurosci. Res. 75*, 162–171.

[118] Petkova, A. T., Ishii, Y., Balbach, J. J., Antzutkin, O. N., Leapman, R. D., Delaglio, F. & Tycko, R. (2002) *Proc. Natl. Acad. Sci. USA. 99*(26), 16742–16747.

[119] Torok, M., Milton, S., Kayed, R., Wu, P., McIntire, T., Glabe, C. & Langen, R. (2002) *J. Biol. Chem. 277*(43), 40810–40815.

[120] Lopez de la Paz, M. & Serrano, L. (2004) *Proc. Natl. Acad. Sci. USA. 101*, 87–92.

Chapter 2

Fragment-Based High Throughput Docking

(Chapter of the book: "Virtual Screening in Drug Discovery", pp 349-378, 2005)

14 Fragment-Based High Throughput Docking

Peter Kolb, Marco Cecchini, Danzhi Huang, and Amedeo Caflisch

14.1 INTRODUCTION

Structural genomics programs around the world are delivering an abundance of three-dimensional (3-D) structures of proteins, some of which are pharmacologically highly relevant. Hence, computer programs for automatic docking of libraries of compounds are being developed further and applied to design drugs against a plethora of diseases including AIDS, Alzheimer's disease, cancer, malaria, and sleeping sickness. In this chapter, we first review the most common approaches for structure-based flexible ligand docking. Some technical improvements for more efficient sampling and more appropriate scoring functions are then presented. Finally, a number of practical suggestions are given for high throughput docking (HTD) with special emphasis on our fragment-based approach.

14.2 OVERVIEW

The basic strategy of any docking approach is to generate a conformation of a putative ligand, which is then placed (or *docked*) in the binding site of a protein target (also referred to as *receptor*). The result of these two operations is usually called a *pose*. A *score* has to be assigned to each pose, thus producing a *ranking*, with the correct pose (i.e., the natural binding mode) at the first rank or at least as close as possible to it.

14.2.1 Defining the Binding Site

Prior to any attempt of docking, the approximate location of the binding site needs to be defined. It is easiest for the case in which the crystal structures of the receptor in complex with some ligands are already known. Usually, the binding site is then defined as the residues lying within a certain cutoff from the ligands.

A greater challenge is presented when only the 3-D structure of the protein is known. In that case, profound knowledge of the function of the protein is necessary. There are programs that analyze the protein surface and provide quantitative information on it, among them GRASP (Graphical Representation and Analysis of Structural Properties) [1] and HYDROMAP [2], which calculate

the electrostatic potential and hydrophobicity map, respectively. Alternatively, some programs use so-called "flood filling" algorithms that attempt to identify cavities on the protein surface. Basically, they fill the space that is not occupied by the protein with points and then roll a large "eraser" over the surface of the protein. All remaining points are said to be in protein pockets [3].

In general, the residues in the binding site are important because their interaction with the ligand is stronger and usually treated in more detail. The binding site residues are explicitly used during the computation of the score and they are sometimes also considered as entities providing anchor points for the positioning of a conformation. Therefore, they should be chosen according to the type and function of the receptor, as well as the program's strategy to determine ligand poses.

Recently, the program AutoDock [4,5] was tested on "blind" docking, that is without defining any selected portion of the protein as binding site [6]. Docking was successful for ligands with less than 10 rotatable bonds, but only at high computational cost (in the order of days). Hence, the definition of the binding site is necessary for virtual screening (VS) of large databases.

Another aspect is the selection of an appropriate protein (and thus binding site) conformation. McGovern and Shoichet have performed a comparative study [7], using the x-ray structures of the complexed and uncomplexed protein as well as conformations obtained by homology modeling of 10 different proteins. The highest enrichment of known ligands in a database was in most cases achieved with the complexed structure. Using a conformation from a complex introduces a bias toward known inhibitors, however, and should thus be complemented by other protein structures in a screening project.

14.2.2 Generating a Pose

Two main types of approaches to obtain a ligand pose have to be distinguished: the ones that use only the complete structure of the ligand and those that follow an incremental strategy. Section 14.2.2.1 and Section 14.2.2.2 refer to the first type; the incremental methods are described in the Section 14.2.2.3.

14.2.2.1 Generation of Ligand Conformations

Typically, docking programs modify only the torsional degrees of freedom of rotatable bonds to produce different ligand conformations. It is important to at least modify the torsional angles of groups carrying hydrogen bond donors (HDO) to allow optimization of this type of interaction. Torsional angles of bonds in rings, double or triple bonds, or single bonds to symmetrical groups (like methyl) are normally kept fixed. In one study with the focus on protein flexibility, the backbone of peptidic inhibitors was considered as being rigid and only "sidechain" flexibility was allowed [5]. A rigorous test of a docking program should consider full flexibility, however [8,9]. An important exception is the docking of small fragments (like benzene or benzamidine), for whom the rigid body approximation is an appropriate description of their limited flexibility [10,11]. Some programs do not

allow ligand flexibility, but the success rates in these cases are low if one does not use the conformation found in the crystal structure [12]. Clearly, such methods can hardly be used to predict the binding modes of "new" ligands. The program DOCK [13,14] also started as a rigid-body docking tool, but ligand flexibility was introduced in DOCK 4.0, using an exhaustive search and conformational refinement with the simplex method [15].

There are two common approaches for generating different ligand conformations:

1. In procedures that search the conformational space of the ligand outside of the binding site, a pool of relevant conformations with low internal energy is generated, and they are subsequently docked rigidly. The sampling of the ligand conformational space can be done exhaustively, modifying each torsional angle in discrete steps [16,17]. Alternatively, the procedure can employ rotamer libraries which assign the most probable values to torsions depending on the atom types [9,15,18].
2. The conformations can be subject to an optimization algorithm, where the torsional angles correspond to the variables of the optimizer. One can further distinguish between two optimizer types: Monte Carlo (MC) searches [3] (also used for *de novo* design by DeWitte et al. [19,20]) and genetic algorithms and other evolutionary approaches [4,8,21–23]. MC approaches use a single conformation that is randomly perturbed and improved. Genetic algorithms (GAs) employ a multitude of information-containing chromosomes (usually referred to as the *population*), which interact with each other and evolve to better solutions. These algorithms are more promising for docking [4], because the energy surfaces to be searched are rugged. MC methods tend to be rather slow, which is a disadvantage for large-scale library screening. Furthermore, if one uses MC–simulated annealing approaches, the additional problem of choosing an appropriate initial temperature and a cooling schedule arises.

14.2.2.2 Defining Ligand Positions

There are several strategies to position and orient the ligand in the binding site:

- The translational degrees of freedom can be encoded in an optimizer.
- The position can be determined by matching the shape of the ligand to the binding site.
- The conformation can be superimposed on a set of points that contain information about the binding site (for references see below).

As an example of approaches that follow the first strategy, the chromosomes in a GA can additionally carry genes for the translational degrees of freedom of the ligand and three (in the case of Euler angles) or four (when quarternions are used [4,24]) variables specifying the ligand orientation.

In approaches that follow the second strategy, the surface of the binding site is compared to the solvent accessible surface of the current ligand conformation. An optimal position is found based on some measure of similarity between those two. LigandFit [3] uses an algorithm developed by Oldfield [25,26], which treats both the binding site and the ligand as a collection of grid points. The shape of such a collection is characterized by a matrix. From the eigenvalues of these matrices, the shape discrepancy can be computed and used to assign a score to each conformation. FRED [17] employs a bump map, which is a Boolean grid representing the receptor, with true values where ligand atoms can potentially be placed. After this initial filtering step, several other scoring functions can be applied, among them Gaussian shape fitting. This function has favorable values when the ligand and the protein have high surface contact and little volume overlap.

DOCK [14,15] follows the third strategy by first filling the binding site with spheres of different sizes. The centers of these spheres are considered as anchors for atoms of the ligand. Variations of this approach at different levels of sophistication include the use of HDOs and HACs (hydrogen bond acceptors) as well as hydrophobic surface points as anchors [27,28]. An example of this is SEED, which was developed to dock small molecules with solvation [10,11]. It uses anchors on the surface of the receptor and performs an exhaustive search on a discrete space by matching donor and acceptor vectors (or vectors of hydrophobic interaction centers) and rotating the ligand around these axes. Other programs use information from the placement of predefined small molecular fragments to match their positions to similar entities in the ligand [16]. The Fragment-based Flexible Ligand Docking (FFLD) program utilizes the results from the docking of small and mainly rigid molecules that have been specifically chosen to match chemical moieties actually present in the ligand [8]. The underlying assumption for all these methods is that the interaction between a protein and a ligand is dominated by some key groups of the ligand. Hence, if the positions of these groups are determined correctly, the rest of the ligand will almost inevitably assume the correct pose.

14.2.2.3 Incremental Methods

Programs like FlexE [9] (an advanced version of FlexX [18]), SLIDE [28], or DOCK 4.0 [15] also try to optimize the interactions of the key groups, but do this individually for each group. The ligand is first split into several units (fragments), the first of which is placed as a seed. Usually, the determination of the pose of the first fragment is done with high accuracy. Sequentially, all the other fragments are connected in their due order, whereby each position is optimized, often exhaustively. At every step, the highest ranking solutions are retained and the next fragment is connected to each of them. It is important to carefully select only a small number of candidate solutions at every step (pruning) to control the exponential increase of possible solutions.

14.2.3 Ranking the Poses

At the beginning of this chapter, we distinguished between exhaustive searches and optimization techniques. The latter minimize an objective function that is usually computationally not too expensive, because it has to be called quite frequently, and a force-field-based binding energy is evaluated for the final ranking. Exhaustive searches use only one energy function.

14.2.3.1 Objective Function

The objective function approximates the interaction energy between ligand and receptor and the internal strains of the ligand and the protein, if the latter is also flexible. Typical components are the intermolecular van der Waals (vdW) and Coulombic energy, and sometimes a term for hydrogen bonds. The internal strain is usually estimated by the intraligand vdW energy and sometimes the dihedral energy. Most objective functions do not take into account terms for bond, angle, and torsional strains. It has been proposed to increase the chances of the optimizer by smoothing the energy landscape. Whitfield et al. [29] introduced a gravitational force that dominates all other forces in the initial steps of the search and then decreases over time. It is assumed that the position of the global optimum does not change due to the smoothing and that only the well depth is modified. Hansmann and Wille [30] developed energy landscape paving, which penalizes scores that are found repetitively. Searches can thus escape local minima and go into regions of different energy.

Most of the docking programs that use physics-based functions (like DOCK [13–15], AutoDock [4,5], and FFLD [8]) employ a grid-based approach for efficiency reasons. These grids contain the Coulombic potential and vdW potential of the protein and avoid the need for recalculating the full energy for every pose during a database screen. Trilinear interpolation [31] is often used to compute the interaction energies from the grid values of the potential.

Empirical-based functions (such as the one used in FlexX [18] and FlexE [9]) use additive approximations to estimate the binding free energy. They contain several terms corresponding to hydrogen bonding, hydrophobic interactions, entropic changes, and sometimes, interactions with metal ions. The coefficients of each term in the sum are obtained from a fit to known experimental binding energies for various protein–ligand complexes [32,33].

14.2.3.2 Binding Energy Function and Postprocessing

After a docking run, the best poses of the ligand can be reranked using a more accurate force field [34,35]. This often contains the same terms as the objective function, but takes longer ranging interactions and ligand and receptor desolvation into account. Sometimes, the ligand pose is also minimized within the receptor using a molecular mechanics force field [36,37]. In our group, ligand poses are normally minimized with CHARMm [36] using the CHARMm22 force field (Accelrys, Inc.), and often also with the TAFF-force-field (Tripos). Additionally, the score and rank of each pose can be redetermined using more accurate energy

functions that include electrostatic solvation like the one in SEED [10,11] or knowledge-based interaction fields like SuperStar [38], potential of mean force (PMF) [39], Small Molecule Growth (SMoG) [40], and DrugScore [41]. The energy rankings produced by the different scoring functions are usually compared, as a number of studies suggest that consensus scoring improves the chance of finding a true hit [42,43].

14.2.3.3 Solvation

The effects of solvation play a key role in molecular recognition events. To calculate the electrostatic contribution to solvation in the continuum dielectric approximation, one could solve the finite-difference Poisson–Boltzmann (PB) equation [44–47] for every new position of the ligand molecule. Considering the current computer power, this would be forbiddingly expensive, especially for HTS. Therefore, only a few docking programs take into account electrostatic solvation effects. The continuum dielectric approximation and the generalized Born (GB) approach [48,49] are used in SEED [10,11], Program to Engineer Peptides (PEP) [50,51], and DOCK [52]. Fairly recently, Arora and Bashford have presented a modified GB approach that estimates desolvation by an integral over the occluded volume [53].

Some docking programs treat solvation effects just with respect to the presence or absence of conserved water molecules that form interactions that are either essential for the protein conformation or necessary to mediate interactions between ligand and protein. Clearly, this approximation completely neglects the bulk properties of water (e.g., dielectric screening). Österberg et al. use grids that have been derived by averaging over several crystal structures, some of which can contain water molecules [5]. Although the method has mainly been developed to incorporate protein flexibility, heterogeneities in the presence of water molecules can be taken into account as well. Schnecke et al. consider water explicitly and have a term penalizing the replacement of water molecules by a hydrophobic group of the ligand [28]. Finally, Rarey et al. have described a method to precompute positions of water molecules and place them if they can form hydrogen bonds with the (partial) ligand during the incremental construction in FlexX [54].

14.2.4 Protein Flexibility

In principle, it would be ideal to allow full flexibility for the protein to model large displacements upon ligand binding. Such studies have already been undertaken [55], but because the computational time was in the order of days for a single ligand, this can clearly not be applied to the screening of large libraries of compounds. As a consequence, flexibility of the protein, if any, is mostly limited to the binding site and its vicinity. Three different approaches shall be highlighted here.

AutoDock [5] incorporates both protein mobility and structural water heterogeneity. It first generates the energy grids for a number of different protein

structures. The program then offers several ways to combine these grids into a single grid. It either computes simple point-by-point averages or weights the different grid points according to their energies and physico-chemical characteristics. This mean grid approach has the advantage that one can still dock to one rigid structure, which facilitates the analysis of the results compared to docking to several distinct conformations. On the other hand, it can only be used to approximate minor displacements. Moreover, the mean grid structure is the product of an averaging scheme and thus might not be observable in reality. Another drawback is the fact that no protein structure is present, but only its representation as a grid. One could thus not follow a multiple step approach (See Section 14.3.4) and do minimization with CHARMm [36], for example.

FlexE [9] is based on a so-called united protein description [56], which is derived from superimposing the backbones of an ensemble of different crystal structures. Variations of the structure in the binding site region are either maintained as distinct possibilities or are combined to one structure in case they are similar. During the incremental construction algorithm, the ligand is placed fragment by fragment into the active site of the united protein description. After each construction step, all possible interactions between the (partially) placed ligand and all instances of the united protein description are determined. The score is then assigned for the (partial) ligand in the best instance.

SLIDE [28] goes one step further and first docks a rigid scaffold into a rigid binding site. Gradually, the other parts of the ligand are attached to the scaffold. Clashes between the ligand and the protein are resolved by allowing rotations of bonds (both in the ligand and the protein) that have been defined as flexible beforehand. The bonds that should be rotated are determined with mean-field theory, which is capable of finding the minimum amount of rotations necessary to resolve all clashes [57–59]. Although flexibility is limited to the binding site residues, this approach comes close to an induced fit.

One of the most thorough approaches besides [55] has been undertaken by Lin et al. [60]. For their relaxed complex method, first long molecular dynamics (MD) simulations of 2 ns were conducted, with snapshots taken every 10 picoseconds (ps). Two candidate compounds were then docked to the ensemble of MD conformations. This technique recognizes the fact that ligands may bind tightly to conformations that appear only infrequently in the dynamics of a protein. However, every molecule has to be docked to a large number of different protein structure which strongly limits the size of the library.

14.3 TECHNICAL IMPROVEMENTS

14.3.1 CURRENT LIMITATIONS

As mentioned above, docking approaches can be described as a combination of two components—the search strategy and the scoring function. Because in most cases the objective function (See Section 14.2.3.1) is also used as the binding energy function (See Section 14.2.3.2), in the following, the term *scoring function*

will be employed. The critical element of the search procedure is the amount of time required to effectively sample the relevant conformational space. The scoring function has to be fast enough to allow its application to a large number of potential solutions and, in principle, be able to effectively distinguish the experimentally observed binding mode from all others explored in the search. Consequently, the scoring function should include and appropriately weight just the energetic contributions that are relevant in the binding process. Nevertheless, an accurate scoring function will generally be computationally expensive and so the function's complexity is often reduced at the expense of a loss in accuracy.

The proper combination of an effective search algorithm and an adequate scoring function, whose global minimum corresponds to the biologically relevant complex, will solve the docking problem in a reasonable amount of time. However, because the approaches published up to date can fail, especially in cross-docking, this ideal combination has obviously not been found yet. Therefore, improvements in the efficiency of the search strategy and the accuracy of the scoring function are required as they will increase the reliability of the docking predictions and reduce the computational requirements, which is important for screening large libraries.

Docking predictions are still prone to fail and often the proposed binding modes do not reproduce the crystal structure of the protein–ligand complex [6,9,35,61]. In case of failure, the predicted binding mode can have a worse or a better score than the x-ray structure of the ligand. In the first case, the search strategy adopted in the docking approach could have been not effective enough. The search algorithm was thus not able to generate a pose sufficiently close to the experimental binding mode. In the second case, the failure might arise from an inadequate scoring function that allows more favorable binding modes than the one in the crystal structure. In the first case, one should focus on the improvement of the search procedure; in the second case, one should concentrate on the optimization of the scoring function.

Unfortunately the situation is much more complicated because the components of a docking protocol are not separate entities and as such they should be improved together. In the first scenario, for example, the scoring function could have played an important role because the resulting energy landscape was not smooth enough to allow the search to proceed efficiently while avoiding premature convergence. Although the scoring function described the protein–ligand interactions well, it was not suitable for the applied search strategy. In the second scenario, it could have happened that the experimentally determined structure was not close to a minimum of the scoring function. In this case, any energy comparison is much less meaningful. Although a proper combination of an efficient search algorithm and an accurate scoring function are the keys for a successful docking protocol, it is certainly not clear what "proper," "efficient," and "accurate" mean. In Section 14.3.2. and Section 14.3.3, we describe some important requirements for both the search strategy and the scoring function and how they are embedded in our docking approach.

14.3.2 Search Strategy

Docking procedures belong to the category of global optimization techniques where the aim is finding the global minimum of the scoring function. A rigorous search algorithm would exhaustively investigate all possible binding modes between the ligand and the receptor. The degrees of translational and rotational freedom of the ligand would be explored along with the internal conformational degrees of freedom of both the ligand and the receptor. However, this is impractical because of the size of the search space, even when considering a rigid protein. Only a small amount of the total conformational space can be sampled and a balance must be reached between the computational expense and the amount of search space examined. A wide range of global optimization algorithms are currently available, but not all of them are suitable for docking. Most optimization algorithms for docking fall into one of three classes—gradient-based algorithms, combinatorial algorithms, and stochastic algorithms [62].

The strength of gradient-based methods is that they efficiently find a local minimum close to the initial conformation. Because gradient-based methods do not allow the system to escape from local minima they have to be combined with other search strategies, such as cycles of MC perturbations and gradient minimizations [63]. Moreover, most scoring functions do not have an analytical gradient.

Combinatorial algorithms have the potential advantage of being extremely fast and effective. The most successful combinatorial algorithms used for molecular docking [10,11,18,64,65] have set themselves apart in their ability to dock libraries of small molecules in a reasonable amount of time. Unfortunately, increasing the number of conformational degrees of freedom leads to an explosion of the dimension of the search space. To be able to sample such large spaces, the computational expense is usually controlled by a discretization of the space, which can restrict the effectiveness of the algorithm.

Stochastic algorithms have the advantage that, irrespective of the dimensionality of the problem and given enough time, they get arbitrarily close to the global minimum. On the other hand, they have the disadvantage that they require a large amount of central processing unit (CPU) time to achieve an acceptable degree of reliability [4,62]. Although computationally expensive, stochastic optimization algorithms seem to be the most suitable for flexible docking. In fact, the dimensionality of the search space and the ruggedness of the binding energy landscape make both gradient-based and combinatorial methods less effective. GAs are stochastic optimization methods that mimic the process of natural evolution by manipulating a population of data structures called chromosomes [66,67]. Although requiring rather large amounts of CPU time, GAs have been shown to effectively explore rough energy surfaces and to be suitable as search strategies for docking [4,8,21–23,68,69]. A GA was chosen as the search strategy for the original version of FFLD [8], the docking protocol developed in our group. During the FFLD evolution, a loop over generations is performed until the maximum number of steps is reached. Starting from an initial random population of chromosomes containing the dihedral angles of

the ligand as genes, the GA repeatedly applies two mutually exclusive evolutionary operators—one-point crossover and mutation. This yields new chromosomes (children) that replace appropriate members (parents) of the population. These nonlinear genetic operators help to overcome the barriers of the binding energy landscape and the search can proceed efficiently. Throughout the simulation, a constant evolutionary pressure is kept by selecting parent chromosomes with a bias toward the fittest. This pressure moves the population toward conformations related to the global minimum and increases the fitness of the individuals. The selection of the members of the population that should be replaced by new chromosomes is a crucial step. To avoid premature convergence, it is important to keep structural diversity. In the search strategy used in FFLD [8], both the energy difference and the conformational similarity are taken into account to determine if a given member of the population should be replaced by a new chromosome. At the end of each GA step, every new chromosome is compared with the old population by the following procedure: if a similar chromosome is found in the old population, it is replaced by the new chromosome only if the energy of the new one is more favorable; otherwise, the new chromosome is discarded. The similarity test significantly improves the efficiency of the search strategy and avoids premature convergence [50].

Following a comparative study of several search engines in AutoDock [4], a hybrid search procedure was introduced in the latest version of FFLD [35]. The hybrid search combines a global optimization procedure based on a GA with a local minimization algorithm to improve exploration of regions within energy basins. Local optimization has been shown to dramatically improve the success rate of the GA search without any loss in efficiency [4]. For the best 10% of the new individuals, a local optimization is performed to improve the ligand fitness before performing the similarity test. To evaluate the performance of the hybrid search procedure implemented in FFLD, it was compared with the GA of the original version [8]. The simulations showed that the hybrid search is more efficient than the canonical GA as it always reached a conformation with lower energy. The results of two docking experiments carried out with both search methods are presented in Figure 14.1. The first experiment, in which a ligand with 10 rotatable bonds was docked in human-immunodeficiency virus type 1 (HIV-1) protease (Figure 14.1, top), shows that the hybrid search procedure is more efficient than the genetic algorithm especially at the beginning of the simulation where the energy gap is large. At about 60% of the evolution the gap decreases and the performance of the two methods is comparable. Docking a ligand with 21 rotatable bonds in HIV-1 protease (Figure 14.1, bottom) shows that the hybrid search procedure performs better during the entire simulation and the energy gap increases until the end. Moreover, the standard deviation of the hybrid search evolutions (shown as error bars in Figure 14.1, bottom) is larger, indicating that it is less prone to converge prematurely. This comparison shows that the local search improves the quality of the docking predictions in case the conformational space of the ligand is large. This is mainly due to the fact that the random perturbations of binary strings performed by the GA during the evolution correspond to radical jumps in the energy landscape and may be too large. On the contrary, the local optimizer is able to refine the large perturbations due to crossover

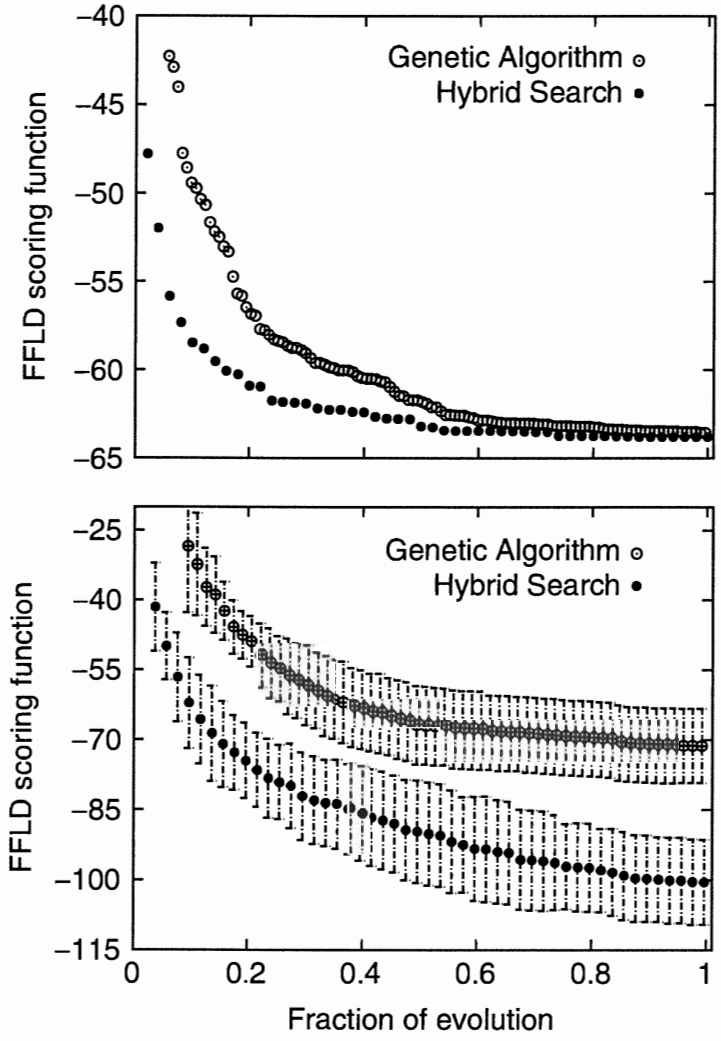


FIGURE 14.1 Evolution of the best individual of the population averaged over 10 docking runs for two different experiments. Empty and filled bullets indicate evolutions performed by GA and hybrid search procedure, respectively. Docking of HIV-1 protease ligands with 10 and 21 rotatable bonds are shown from top to bottom, respectively. In the bottom plot, the vertical bars show the standard deviation computed over 10 docking runs.

and mutations and leads to a better investigation of the energy landscape. The results of this docking study [35] suggests, in agreement with previous studies [4], that hybrid search methods should be preferred to canonical GAs.

The similarity test and the hybrid search procedure are just examples of possible means one can adopt in a protocol to increase the efficiency and accuracy of the search algorithm. However, the study clearly indicates that there is still room for improvement and that novel concepts can be effective. It is worth stressing again that the search algorithm is only half of the docking problem; the other factor to be incorporated into a successful protocol is the scoring function. In Section 14.3.3, the requirements for a scoring function that are suitable for docking are discussed.

14.3.3 Scoring Function

Underlying any docking approach is a model of ligand–protein interactions describing molecular recognition. In principle, a complete thermodynamic description of this process involves contributions from several balancing factors, including solvent reorganization, conformational entropy, and vdW and electrostatic interaction energies. For biomolecular systems, it is difficult to evaluate these terms with sufficient accuracy to permit quantitative predictions. Moreover, the complete energy function necessary for prediction of accurate binding affinities may not be suitable for docking simulations. The scoring function used in docking simulations should be a simple model of ligand–protein interactions rather than an estimation of the free energy of binding. It must be simple enough to permit a rapid evaluation and, more importantly, the resulting energy landscape must be smooth enough to allow the search to proceed efficiently without getting trapped in local minima. Nevertheless, a scoring function that is suitable for docking needs to be accurate, because it must be able to distinguish the experimental binding mode from all the other modes explored by the search algorithm.

With respect to this point, Verkhivker et al. [69] suggested that such an energy function should fulfill both a thermodynamic and a kinetic requirement. In other words, the energy related to the crystallographic structure of the ligand in the complex must be the global minimum of the binding energy landscape (*thermodynamic requirement*), but at the same time this conformation must be accessible during the search (*kinetic requirement*). The complexity of a complete and accurate force field that describes the binding process precisely, although it would fulfill the thermodynamic requirement, typically results in a rugged energy landscape and thus does not meet the kinetic criterion of the docking problem. The multitude of energetically similar but structurally different local minima inevitably leads to kinetic bottlenecks that dramatically reduce the frequency of successful structure predictions. This is the case for standard molecular mechanics force fields [36,37], because they have not been designed to reduce the ruggedness of the energy landscape. One of the critical factors that determines the success rate in predicting the structure of ligand–protein complexes is the roughness of the binding energy landscape [68,69]. Consequently, the applicability of standard force fields in docking is limited and simpler molecular recognition models that fulfill both the thermodynamic and kinetic requirements are to be designed and developed.

A fundamental component of models for molecular recognition is the steric energy function, which is based on surface complementarity. However, this term alone is not sufficient to distinguish effectively between alternative binding modes. Electrostatic interactions may provide additional specificity to discriminate between true and false solutions and they should be embedded in the scoring function. Finally, an intraligand energy term is also required; it largely reduces the conformational space to be investigated by preventing strained dihedrals and steric clashes among atoms of the ligand. Hence, the three key elements of a scoring function necessary for robust structural assessment during docking are:

1. Ligand–protein steric interactions
2. A simple description of ligand–protein electrostatics
3. An intraligand strain

In the FFLD docking approach developed in our group [8], the scoring function is

$$\Delta E_{total} = E_{dihedral}^{ligand} + E_{vdW}^{ligand} + E_{vdW}^{inter} + E_{polar}^{inter} \tag{14.1}$$

The dihedral energy of the ligand $\left(E_{dihedral}^{ligand}\right)$ has recently been implemented in FFLD (D. Huang, unpublished results) using the lowest order terms of a cosine expansion for each torsion. The second $\left(E_{vdW}^{ligand}\right)$ and the third $\left(E_{vdW}^{inter}\right)$ terms of Equation 14.1 are intraligand and ligand–receptor vdW energies, respectively. Both terms are described as the sum of an attractive dispersion and a steep repulsion term by the 6-12 Lennard-Jones potential. The last term in Equation 14.1 is the protein–ligand polar interaction energy $\left(E_{polar}^{inter}\right)$. The intermolecular polar term approximates electrostatic interactions and includes hydrogen bonds (HB) and unfavorable polar contacts (UP), namely two HAC (or HDO) atoms close to each other. Hence

$$E_{polar}^{inter} = \sum_{i=1}^{N_{HB}} E_i^{HB} + \sum_{i=1}^{N_{UP}} E_i^{UP} \tag{14.2}$$

where N_{HB} and N_{UP} are the number of hydrogen bonds and the number of unfavorable polar contacts, respectively. The energies E_{HB} and E_{UP} are approximated by constant values [35]. Distance- and angle-dependent criteria are considered for the definition of a hydrogen bond, but only a distance dependence is applied for unfavorable polar contacts. Originally, the distance dependence of both terms in Equation 14.2 and the directionality of the hydrogen bonds follow simple step functions (Figure 14.2, top left and top right, dashed lines) that are efficiently evaluated [8]. The steep repulsive part of the Lennard-Jones potential directly affects the height of the energy barriers and generates a rough energy surface. To reduce the steepness of this energy component, an intermolecular soft-core vdW term was implemented [8]. Following previous studies by Gehlhaar et al. [68], the repulsive part of the Lennard-Jones potential was linearized in FFLD, such that the functional form has a finite value when the interatomic distance approaches zero (Figure 14.2, bottom). The intermolecular soft-core vdW does not penalize binding modes with small atomic interpenetrations of the ligand with the protein and permits the formation of unphysical states that could open multiple pathways leading to the crystal structure. These states, otherwise forbidden by the presence of realistic energy barriers in standard force fields, may provide kinetically accessible routes to the global minimum.

In a recent study [35], a significant improvement with respect to the original version of our docking approach [8] has been observed by replacing the step

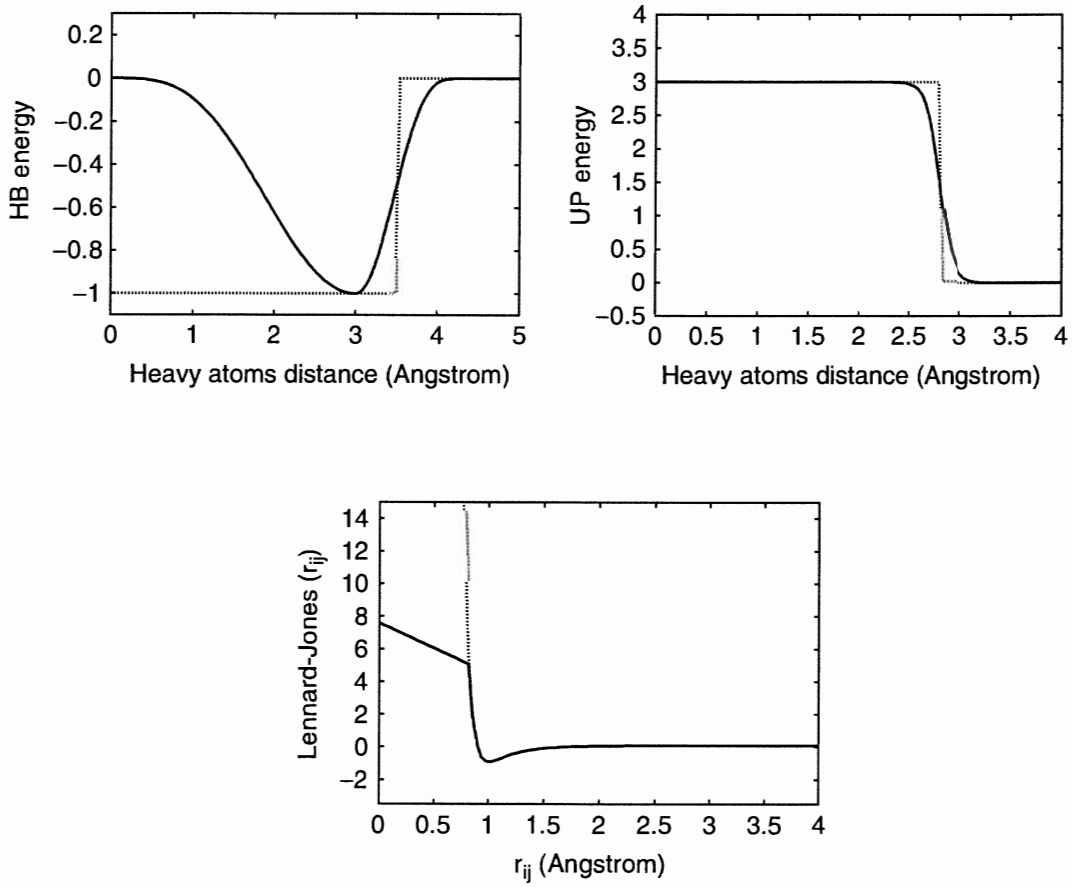


FIGURE 14.2 The distance dependence of hydrogen bonds (HB), unfavorable polar contacts (UP) and ligand-receptor vdW interactions is presented from left to right, respectively. The smooth functions (solid lines) [35] used for replacing the original stepwise functions (dashed lines) in the intermolecular polar interaction term are shown. On the bottom, the intermolecular soft-core vdW (solid line) [8] is compared with the 6-12 Lennard-Jones potential (dashed line). Values are in kcal/mol.

functions in the ligand–receptor polar interaction term $\left(E_{polar}^{inter}\right)$ with *smooth* functions. Smooth functions allow the optimization of the hydrogen bonding pattern avoiding discontinuities on the energy landscape. The continuous gradient can guide the search algorithm toward lower energy conformations at every point. In the latest version of the FFLD docking program, a sigmoidal function was used to describe the unfavorable polar contacts and bathtub-shaped functions were used for the distance dependence and the directionality of the hydrogen bonds (Figure 14.2, top left and top right, solid lines). Furthermore, it was observed that the distance- and angle-dependence in the polar term significantly reduced the noise arising from the energy degeneration of structurally different ligand conformations and improved the convergence of the docking runs [35].

Previous works by Gehlhaar and Verkhivker [68,69] suggested that a dynamical modification of the scoring function is helpful. In their docking experiments, an adaptive scoring function based on a piecewise linear potential was used. During docking, the height of the energy barriers had been continuously augmented by

increasing the repulsive term of the potential. Thus, in the later stages of the simulation, this adaptive procedure narrowed the search to only a few energetically favorable binding modes, funneling the algorithm to the global minimum. According to the authors, the adaptive softness of the energy function facilitated the conformational search both by promoting escape from local minima and by destabilizing alternative solutions. Increasing the repulsive term of the potential yields a rougher energy landscape, but the energy function becomes more and more accurate and leads the search to the global minimum. Similar dynamical modifications of the energy function have been adopted to mimic the docking funnel [29,30,55]. Although the essential idea is rather simple, no general rules for adapting the potential are available and the optimal way for scaling the barriers may be strictly dependent on the system explored. Moreover, if the scaling is not accomplished in a proper way, the adaptive scoring function might not fulfill the kinetic requirement. Because of these limitations, we and others [35,70–73] have chosen an alternative approach. This is described in Section 14.3.4.

14.3.4 Multiple-Step Docking

Combining different scoring schemes into a single docking approach is a useful method to increase the effectiveness of a docking protocol. A two-step strategy makes use of a simple molecular recognition model based on the minimal frustration principle [68,69], followed by a more accurate energy evaluation to rank the docking solutions. When using multiple-step procedures, there is a clear distinction between the objective function, which is fast but approximative, and the binding energy function (See Section 14.2.3.1 and Section 14.2.3.2).

The basic assumption behind multiple-step approaches is that there is at least one low-lying minimum of the objective function inside the global minimum basin of the binding energy. The fast objective function is then thought of as a coarse-grained description of the more accurate binding energy function. The first step intends to overcome the kinetic bottlenecks of the accurate energy function by using a simpler and much less frustrated energy model. After the first step of the procedure, the final set of ligand conformations can undergo a gradient-based minimization with a standard force field. The minimized conformations are then ranked according to their energy. Multiple-step docking approaches are widely used and have been published [70–73]. A multiple-step procedure was also applied in the most recent version of our docking approach [35]. The results of FFLD [8] were postprocessed by CHARMm minimization [36] of the flexible ligand in the rigid receptor. The docking study showed the effectiveness of a multiple-step strategy. It was possible to correctly reproduce the binding mode of highly flexible inhibitors (up to 22 rotatable bonds) of HIV-1 protease, if the strain in their covalent geometry upon binding was not too large. Moreover, it was observed that the postprocessing step led to more reliable predictions and improved the success rate of the docking experiments [35].

14.4 PROTOCOLS

In this section, we will explain the use of our docking approach. However, many of the guidelines and recommendations introduced here will also hold true when using other docking programs.

14.4.1 Our Docking Approach

The SEED/FFLD approach uses a GA to optimize ligand conformations and previously docked fragments to place the ligand in the binding site. It relies on the assumption that the most significant interactions with the protein are formed by three or more fragments of the ligand. Hence, it should be possible to first investigate the binding modes of the fragments and then use this information to place the whole molecule. This docking approach consists of four separate steps, the principles of which shall be described below. A more detailed protocol can be found in the following subsections and the original articles [8,10,11,35].

At first, those parts of the ligand that are supposed to account for most of the interactions (the fragments, Figure 14.3) have to be defined. This choice is rather important, for example, fragments that are too small will yield anchor positions that cannot discriminate the physicochemical characteristics of the binding site. A computer program has been developed to automatically choose at least three fragments (P. Kolb et al., unpublished), because the matching algorithm employed in the last step uses triangles. In the second step, the selected fragments are minimized with a force field to obtain low energy conformations. Subsequently, they are docked as rigid molecules with SEED [10,11] (Figure 14.4). As described before, SEED uses polar and hydrophobic vectors as anchors. The polar vectors are distributed around HDOs and HACs, whereas apolar vectors are used to mark hydrophobic regions. The latter are obtained by placing a low dielectric sphere (methane) at equal intervals on the solvent accessible surface of the protein. Points that have a favorable interaction energy are retained and the vectors are defined by joining each point with the corresponding atom center. During docking, every vector is matched to the complementary vectors on the fragments and the fragments are rotated exhaustively around these vector-defined axes. For each fragment position on each SEED point, a binding energy, which

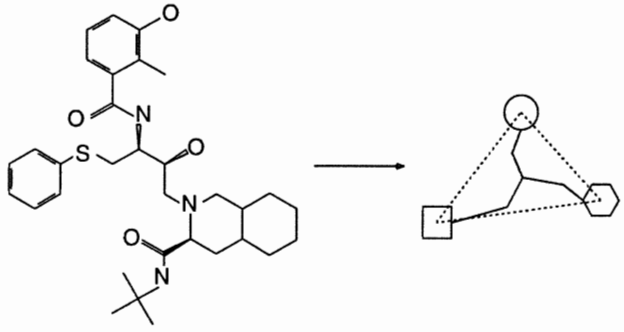


FIGURE 14.3 Schematic depiction of the fragment selection process. The molecule is Viracept (Agouron/Pfizer).

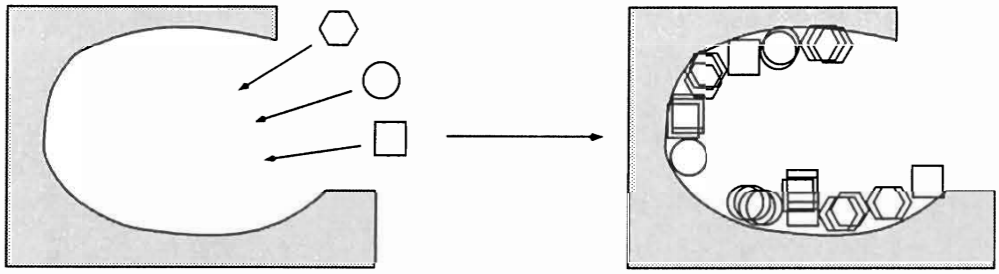


FIGURE 14.4 Schematic depiction of the docking process of the small fragments.

includes electrostatic solvation, is evaluated. Thus, if the fragments chosen are rigid (which is the case for small molecules and aromatic systems), the ranking is determined with high reliability. The information obtained from SEED consists of the 3-D coordinates of the geometrical centers of the fragment poses as well as their binding energies. Each fragment pose is one possible corner point of the placement triangle used in the last step. On average, a SEED run yields up to 100 poses per fragment type.

In the third step, this number is reduced to obtain a manageable number of possible triangle combinations. In practice, we reduce it to 20, using a clustering method which is based both on geometric proximity and the value of the binding energy for each pose [35]. For each fragment, the 20 points define a map that contains the important information from SEED and is still diverse enough to offer useful anchor points (Figure 14.5). Diversity is especially important because using only the top-ranked poses of the fragments does not always lead to the solution. This is due to the fact that the binding mode of the entire ligand is a compromise that tries to satisfy most of the fragments.

The fourth and last step is the docking of the complete putative ligand. This is done with the program FFLD [8], which uses a scoring function consisting of ligand dihedral and vdW energy, and protein–ligand polar and vdW contributions (See Section 14.3.3). Ligand conformations are generated and optimized by a GA, which encodes the torsional angle values of the rotatable bonds. For each conformation, the geometrical centers of the three fragments define a triangle. Based on the side lengths of the ligand triangle, FFLD finds those SEED points that form triangles of approximately the same shape. It then

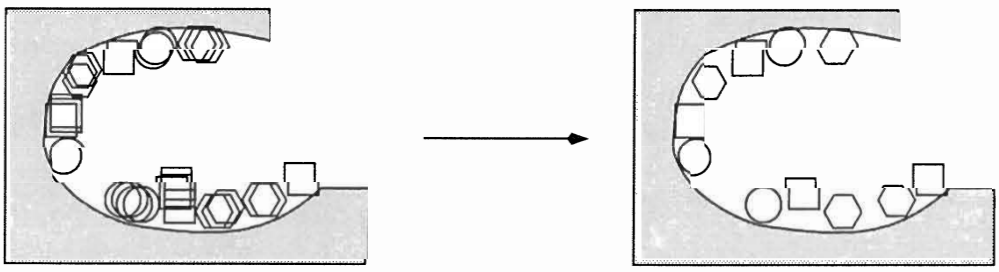


FIGURE 14.5 Schematic depiction of the clustering procedure. The different fragment types are shown for clarity.

tries to match the ligand triangle with each of the possible SEED triangles using a least-squares-fitting method (a variant of the Kabsch algorithm [74]) and assigns the score of the best placement to this conformation (Figure 14.6). The output of FFLD consists of the final poses for all conformations, usually 100 to 200 in total. It is worth noting that, because every conformation yields multiple poses at each step, FFLD will not only find the best binding mode, but also a number of alternative binding modes of comparable score. The alternative binding modes, in fact, can be used as a starting point for further postprocessing with more accurate energy functions.

14.4.2 Preparation of the Library of Compounds

The first and most basic requirement is that the ligand is a chemically complete molecule (i.e., all valences must be satisfied). Special care must be taken to specify the correct bond types, because this will be the basis for the definition of the bonds that are rotatable. Another main concern is the correct assignment of the partial charges. These are needed for the calculation of the interaction energy in SEED and the postprocessing step. We use the modified partial equalization of orbital electronegativity (MPEOE) method developed by No et al. [75,76] as implemented in WITNOTP (A. Widmer, Novartis Pharma AG, Basel, unpublished), which yields partial charges consistent with those of the protein atoms in the CHARMm22 force field. Other implementations should also give reliable partial charges, but we have not tested them.

As a prerequisite to docking, one has to consider the state of ionizable groups in the protein (see below) and the ligand. Because the physiological conditions for protein–ligand complexes are in most cases close to pH 7, acidic groups are usually in a deprotonated and basic groups in a protonated state. A pK_a calculation could be done with a finite-difference Poisson solver in case of uncertainties. For a heterogenous library of compounds, it is much more difficult to assign formal charges. We usually check for groups where the assignment is evident (e.g., primary, secondary or tertiary amines, which are positively charged). Afterward, atom types for the CHARMm22 force field have to be assigned. Any ligand should furthermore be minimized with an accurate force field to obtain a low-energy conformation.

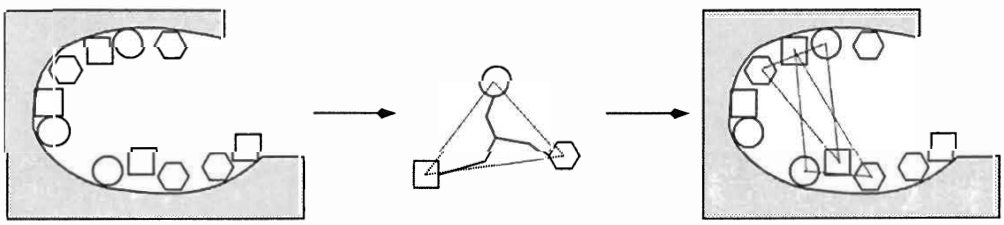


FIGURE 14.6 The docking process: FFLD tries to place the triangle defined by a conformation of the ligand (generated with the GA) on the anchor points computed by SEED.

14.4.3 Fragment Choice

The decomposition of a ligand into fragments and the choice of the anchor fragments have been automatized recently (P. Kolb et al., unpublished). We will list the major rules here as they can be of general interest. The decomposition is guided by the fact that SEED treats all molecules as rigid. Hence, preference is given to aromatic rings and other small rings and molecules that contain several amidic, double, or triple bonds. The fact that nonaromatic ring systems might have several distinct conformations can be accounted for by the ability of SEED to dock multiple (predefined) conformations at the same time. If one of these conformations can be docked with a lower binding energy than the others, it will automatically be chosen in the subsequent steps, because it will receive higher ranks.

The selection follows a few simple rules:

1. All atoms in a fragment must be connected by rigid or terminal bonds (for the definition of rigid bonds see above).
2. Large fragments are preferred because there are more steric constraints for large entities, as a consequence these should be positioned first.
3. Cyclic fragments are preferred because they usually are more rigid than acyclic moieties.
4. Because the fragments should be involved in the most significant interactions, those that contain HDOs and HACs are selected. Charged groups usually do not make such good anchors, because they tend to be positioned at the borders of the binding site, which are more exposed to the solvent. (However, there are exceptions as in the case of thrombin, where a favorable electrostatic interaction is provided by a charged aspartic acid in the specificity pocket [8].
5. Fragments that are close to the center of the molecule are omitted, especially if they have a high number of substituent groups. Such central or scaffold fragments will hardly ever form specific interactions.
6. Finally, fragments should not overlap (i.e., one atom should not be part of two fragments), because this would mean that there are no rotatable bonds in between, so their relative position cannot be changed.

These rules can be exemplified with the molecule XK263 (DuPont Merck, Figure 14.7). In principle, there are three fragment types that could be chosen—naphthalene, benzene, and the cyclic urea in the center. The largest fragment would be the cyclic urea. According to Rule 5, this is not a good choice as it is the core fragment and has four substituents. Furthermore, it is the most flexible of the three types, which is another point against its choice according to Rule 2. The remaining two types are aromatic and thus a recommended choice (Rule 1). Finally, it is better to select the two naphthalenes, because they are larger than the benzenes (Rule 2).

A more difficult choice is presented by acetyl-pepstatin (Figure 14.8), because it has no rings and almost no rigid bonds. All the fragments obtained by the application of Rule 1 are therefore small. All the larger fragments with a rigid

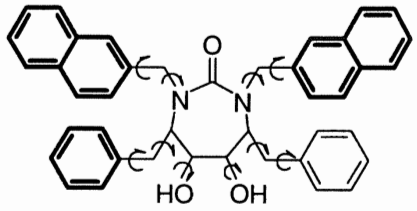


FIGURE 14.7 XK263 (Dupont Merck) is a nanomolar inhibitor of HIV-1 aspartic protease (PDB accession code of the complex: 1HVR). Selected fragments are bold. Curly arrows denote rotatable bonds.

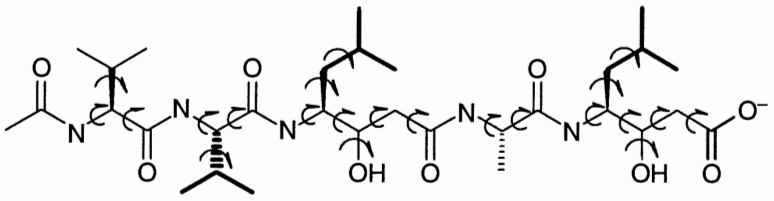


FIGURE 14.8 Acetyl-pepstatin is a micromolar inhibitor of HIV-1 aspartic protease (PDB accession code of the complex: 5HVP). Selected fragments are bold. Curly arrows denote rotatable bonds.

bond (the amide groups) are located in the backbone and will not make good anchors (Rule 5). One of the few choices remaining is to select three i-butanes (the sidechains), which are preferable with respect to the terminal carboxylic group, because this group is charged (Rule 4).

14.4.4 Protein Preparation

It has to be emphasized that the preparation of the protein is a crucial step in the protocol and should be done carefully. It is not advisable to use automatic methods, as they cannot take into account all eventualities and special cases.

14.4.4.1 First Checks

The attention of the experimenter should be turned to all specific and unusual details, like nonstandard amino acids (e.g., cysteine-sulfonic acid, selenomethionine, etc.). Furthermore, the protein can contain prosthetic groups, cofactors, or other small molecules. Prosthetic groups should be kept for the docking run in all cases, because they will most probably be present in the protein in its native environment. Whether or not cofactors should be considered, depends on the system. Most probably, they can be removed, since they will not compete with an inhibitor, unless they have a strong affinity to the protein by themselves and will be present in the binding site most of the time. In general, small molecules (such as polyethylene glycol) are due to the crystallization conditions and should be removed. The final decision, however, has to be taken *ad hoc* for every system.

In any case, one should check in the pdb-file that no atoms are missing in the aforementioned residues and molecules, because most structure manipulation programs do not check on nonstandard residues automatically. Quite frequently,

crystal structures will lack even whole parts of the protein due to poor electron density in disordered regions. This fact is usually commented on in the pdb-file or in the paper. It is then up to the researcher to decide if this is negligible or not. Judging from our experience, in the majority of cases, these incomplete regions are far away from the binding site. Thus, they will not have a great influence on the binding energy evaluation. Unless there are only one or two amino acids missing, it is not advisable to rebuild the protein in those regions. The error introduced by guessing the conformation without proper equilibration will probably be larger than the error due to the absence of the residues.

Another special case are ions. Those that are required for the stability of the protein should be kept, especially if they are close to the binding site. An ion in the binding site should always make a favorable interaction with an oppositely charged group in the ligand. It is advisable to determine the charged warhead for the candidate ligands *a priori* and discuss the simpleness of synthesis of the resulting compounds with a medicinal chemist.

Lastly, the presence of disulfide bonds has to be investigated. Information whether or not there are any should be listed in the pdb-file in a line commencing with "SSBOND." However, it is safer to visualize all cysteine residues. If the sulfur–sulfur distance between two cysteine residues is around 2 Å and the relative geometry is right, they will most likely form a disulfide bond.

14.4.4.2 Charged Residues

Special care should be exercised when treating residues with ionizable groups. The most sophisticated approach is to solve the finite-difference Poisson equation to calculate the pK_a of all titratable groups. If the *in vitro* tests are done at physiological pH, we normally assume both basic and acidic sidechains as well as the terminal carboxyl and amino group as ionized.

The situation for histidine residues is more complicated. First, one has to select a protonation state and then, in the case of monoprotonation, also which nitrogen (δ or ε) should be protonated. To properly assign the protonation state of the histidines, it is important to consider the local environment of these residues in the folded structure of the protein. At low pH ($pH \leq 6$), a diprotonated state should be assigned to histidines partially or fully exposed to the solvent. For calculations at physiological pH, a monoprotonated state is commonly preferred and we assign a monoprotonated state to the histidines irrespective of their position. If the environment does not indicate a clear preference for one of the two variants because of potential HACs or steric hindrance, we arbitrarily choose the δ-protonated variant.

Related to the issue of the charged residues is the choice of the interior dielectric constant of the protein, which is necessary for SEED. The value of this constant influences the strength of the coulombic interactions and can lead to significantly different results, as model calculations have shown (Majeux et al., unpublished results). Previously, values ranging from 1 to 4 have been used [10,11]. It is useful to perform preliminary docking runs with interior dielectric values of 1, 2, and 4 and compare the results with available crystal structures.

14.4.4.3 Adding Hydrogens

It is necessary to add hydrogens, because files in the Protein Data Bank (PDB) usually do not contain any. This should be done with a program like CHARMm [36] using the HBUILD module, which first places those hydrogens whose positions can be determined unambiguously, such as hydrogens connected to a peptidic nitrogen, and afterwards performs exhaustive searches to place hydroxyl hydrogens on serine, threonine, and tyrosine. To assign atom types, we use the atom type definition of the CHARMm22 force field. It has proven useful to recheck on all nonstandard residues to verify the correct assignment. Finally, the hydrogens should be minimized with an appropriate force field while keeping the protein backbone rigid.

14.4.4.4 Binding Site Definition

As mentioned above, this step is of high importance. To begin with, one should have a look at the publication describing the crystal structure and the interactions. The basis for the selection of the residues belonging to the binding site will most often be the pose of a known ligand. If such information is not available, one has to select the binding site by hand. In that case, in-depth knowledge of the function of the protein or crystal structures of closely related proteins of the same family are necessary.

We select the binding site by first determining all protein atoms that are within a cutoff radius of 5 Å from any ligand atom. It is important that there is a clear inside and outside of the binding site to avoid the positioning of anchors in solvent-exposed regions of the protein. Hence, selecting residues whose sidechains point away from the binding site have to be avoided. To achieve this, only residues which have at least 50% of their atoms within the cutoff distance are marked as members of the binding site. The cutoff should not be too small, as the bias toward the binding mode of the known ligand would be too big and no alternative ones could be detected. On the other hand, because the binding site residues are providing the anchor points for SEED, the number of anchors correlates with the number of residues. Thus, docking would take increasingly long as the binding site becomes larger and would additionally yield too many solutions, which are then difficult to rank. If a large binding site is really needed, it is probably better to split it into several (overlapping) sectors. Sometimes, it is advisable to manually alter the definition until one is satisfied with the distribution and the number of the anchor points. In this case, one has to remember that the binding mode (and consequently the ranking of a library of compounds) might be affected by the human intervention, which is usually based on previous knowledge. This bias might preclude interesting surprises like alternative binding modes [77].

As was mentioned before, SEED puts anchor vectors on atoms of the binding site residues. Clearly, only vectors pointing inside the binding site should be used. For that reason, the latest version of SEED employs a cutoff based on the angle between the vector and predefined points in the binding site (usually the

heavy atoms of a native ligand) for choosing the most suitable ones [35]. Using the atoms of a ligand from a known complex to define the binding site does not introduce a bias, though, and corresponds to the situation in an advanced drug design program, where one or more crystal structures of protein/ligand complexes have already been solved.

Another critical issue is the ionization state of groups in the binding site. This is probably best illustrated by the case of the aspartic proteases, which contain an aspartyl dyad in the cleavage site. Piana et al. [78] have shown that, besides the pH, the ligand has an influence, as it can stabilize either the neutral, negatively or dinegatively charged form of the dyad state. Consequently, the charge state of the dyad can influence the types of ligands that will receive a high ranking.

14.4.4.5 Conserved Water Molecules

In many proteins, water molecules located at distinct positions can play a crucial role because they provide important interactions with the ligand. Wrongly positioned water molecules, on the other hand, can impede docking and make the detection of the correct binding mode impossible. Deciding which water molecules to keep is not trivial. Evidence can come from multiple x-ray structures with different ligands. If a water molecule is repeatedly found at the same position and also forms hydrogen bonds with the ligand, it is likely to be conserved because of structural relevance. Additional help is offered by prediction programs such as ConSolv [79], which compares the ligand-free form of the protein with the complex.

Our example, HIV-1 protease, for which numerous x-ray structures are available, normally contains a water molecule bridging the two flaps and the inhibitor. This water is necessary if one wants to reproduce the binding mode of acetyl-pepstatin in its native protein structure, 5HVP. The structure of 1HVR, however, does not contain a water molecule at that specific position. During binding, the carbonyl group of the cyclic urea displaces this water and directly stabilizes the two flaps of the protease. Therefore, docking the ligand XK263 in 1HVR requires the water site to be empty. It is possible to reproduce its binding mode only after removal of the water. However, it is not possible to know this *a priori* for every molecule in a large database for screening. Hence, in the absence of further information, we suggest removing all water molecules from the binding site.

14.4.4.6 Reference Structure

For every new project, the setup of the approach chosen for docking should be validated. The most common way to do this is by redocking a ligand to the corresponding protein structure from the complex. However to judge the performance of the method, it is crucial not to use the exact pose of the ligand from the crystal structure. This pose is the time-average over the ligand poses during the collection of the diffraction data (as is the case for the conformation of the protein).

Thus, it is likely that, according to the parameters of the applied scoring function, some atom positions have clashes with the protein. This problem can be solved by minimizing the ligand within the binding site with a gradient-based method applying the same scoring function as will be used for docking, while keeping the protein rigid. The minimized ligand then offers an appropriate reference structure for redocking calculations.

The ligand conformation which is used as input structure for the docking experiments should have been minimized with a force field outside of the binding site to remove any geometrical bias. However, one has to bear in mind that the force field will not only modify the torsion angles, but also bond lengths and bond angles. If the strain in the ligand conformation is large upon binding, the minimization outside of the receptor might yield a covalent geometry that is not compatible with the binding site. Therefore, because in the docking search only torsional degrees of freedom are considered, the docking approach might not be able to reproduce the experimental binding mode [35].

14.4.5 Running SEED

SEED provides the anchors for the final docking procedure. Thus, it is worth analyzing the SEED results in detail. One should have a close look at the binding site with a molecular viewer to see the distribution of the polar and apolar vectors used by SEED to dock the fragments. If a project is in an advanced stage and a considerable amount of structural information is available, the user should eventually change the number of the polar and apolar vectors as well as the definition of the binding site or the interior dielectric constant.

14.4.6 Running FFLD

The only parameters that should be modified in FFLD are the input values for the hybrid search algorithm. It has to be emphasized that optimal input values depend on the shape of the energy hypersurface and can thus hardly be predicted. As the limiting factor rather is the computer power, the user might want to select fewer chromosomes or fewer steps (which results in fewer energy evaluations) or a smaller frequency for the local search.

It is important, however, to perform multiple runs with different seeds for the random generation of the initial population. As with any stochastic search method, the hybrid search can be trapped in local minima. This is only detectable by comparing the results of many runs, therefore we typically perform 10 runs with different random seed numbers per ligand. Moreover, to judge the quality of the predictions, it is important to have a look at the convergence rate (i.e., which percentage of the different runs reach a similar conformation) [35]. This finding was obtained in a cross-docking study (which corresponds to the situation in a screening project) on 5 complexes of HIV-1 protease. Each of the 5 ligands was docked into all protein structures except its native one, which resulted in a total number of 20 docking experiments. For each docking experiment, convergence toward the lowest energy conformation (which is not

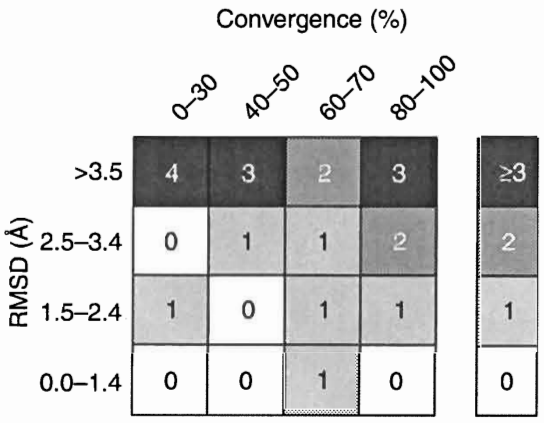


FIGURE 14.9 The density plot with the frequency of a certain rmsd from the experimentally determined structure for a given amount of convergence in 10 GA runs with different seeds. As an example, the "3" in the top right corner means that in 3 of the 20 docking experiments between 8 and 10 runs converged to the same conformation and this conformation has a rmsd larger than 3.5 Å from the experimental structure.

necessarily identical to the experimental structure) in 10 FFLD runs with different seeds was determined. The convergence values were then used to build a density plot that reports the frequency of finding a binding mode with a certain root-mean-square deviation (rmsd) from the experimental structure for a given amount of convergence (Figure 14.9). This density plot is almost upper triangular, which implies that experiments with less than 60% of convergence have probably failed to locate the global minimum. Consequently, these runs should not be relied on. On the other hand, a high convergence is no guarantee for successful docking, as is shown by the high number of runs that fully converged on a wrong structure (Figure 14.9, top right corner). The reason for this is probably to be searched for in the oversimplified nature of the energy function, which precludes an accurate detection of the solution. Taken together, these results suggest that a high convergence rate in multiple GA runs may be a necessary, although not sufficient, criterion for a good prediction.

ACKNOWLEDGMENTS

We thank Dr. Nicolas Majeux for interesting discussions. We also thank Fabian Dey for comments on this chapter. The development of our docking programs has been financially supported by Novartis, Aventis, and the Swiss National Center of Competence in Research (NCCR) in Structural Biology.

REFERENCES

[1] A. Nicholls, K. Sharp, B. Honig, Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons, *Proteins: Structure, Function and Genetics* 11:281–296, 1991.

[2] M. Scarsi, N. Majeux, and A. Caflisch, Hydrophobicity at the surface of proteins, *Proteins: Structure, Function and Genetics* 37:565–575, 1999.

[3] C.M. Venkatachalam, X. Jiang, T. Oldfield, M. Waldman, LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites, *J. Mol. Graphics Modelling* 21:289–307, 2003.

[4] G.M. Morris, D.S. Goodsell, R.S. Halliday, R. Huey, W.E. Hart, R.K. Belew, and A.J. Olson, Automated docking using a Lamarckian Genetic Algorithm and an empirical binding free energy function, *J. Comput. Chem.* 19:1639–1662, 1998.

[5] F. Österberg, G.M. Morris, M.F. Sanner, A.J. Olson, and D.S. Goodsell, Automated docking to multiple target structures: incorporation of protein mobility and structural water heterogeneity in AutoDock, *Proteins: Structure, Function, and Genetics* 46:34–40, 2002.

[6] C. Hetényi, and D. van der Spoel, Efficient docking of peptides without prior knowledge of the binding site, *Protein Sci.* 11:1729–1737, 2002.

[7] S.L. McGovern and B.K. Shoichet, Information decay in molecular docking screens against holo, apo, and modeled conformations of enzymes, *J. Med. Chem.* 46:2895–2907, 2003.

[8] N. Budin, N. Majeux, and A. Caflisch, Fragment-based flexible ligand docking by evolutionary optimization, *Biol. & Chem.* 382:1365–1372, 2001.

[9] H. Claussen, C. Buning, M. Rarey, and T. Lengauer, FlexE: Efficient molecular docking considering protein structure variations, *Algorithmica* 308:377–395, 2001.

[10] N. Majeux, M. Scarsi, J. Apostolakis, C. Ehrhardt, and A. Caflisch, Exhaustive docking of molecular fragments with electrostatic solvation, *Proteins: Structure, Function, and Genetics* 37:88–105, 1999.

[11] N. Majeux, M. Scarsi, and A. Caflisch, Efficient electrostatic solvation model for protein-docking, *Proteins: Structure, Function, and Genetics* 42:256–268, 2001.

[12] A. Fahmy and G. Wagner, TreeDock: a tool for protein docking based on minimizing van der Waals energies, *J. Am. Chem. Soc.* 124:1241–1250, 2002.

[13] E.C. Meng, B.K. Shoichet, and I.D. Kuntz, Automated docking with grid-based energy evaluation, *J. Comput. Chem.* 13:505–524, 1992.

[14] I.D. Kuntz, E.C. Meng, S.J. Oatley, R. Langridge, and T.E. Ferrin, A geometric approach to macromolecule–ligand interactions, *J. Mol. Biol.* 161:269–288, 1982.

[15] T.J.A. Ewing, S. Makino, A.G. Skillman, and I.D. Kuntz, DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases, *J. Computer-Aided Mol. Design* 15:411–428, 2001.

[16] A.N. Jain, Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine, *J. Med. Chem.* 46:499–511, 2003.

[17] OpenEye Software, FRED, 2002. http://www.eyesopen.com/products/applications/fred.html.

[18] M. Rarey, B. Kramer, T. Lengauer, and G. Klebe, A fast flexible docking method using an incremental construction algorithm, *J. Mol. Biol.* 261:470–489, 1996.

[19] R. DeWitte, and E. Shakhnovich, SMoG: *de novo* design method based on simple, fast, and accurate free energy estimates. 1. Methodology and supporting evidence, *J. Am. Chem. Soc.* 118:11733–11744, 1996.

[20] R. DeWitte, A. Ishchenko, and E. Shakhnovich, SMoG: *de novo* design method based on simple, fast, and accurate free energy estimates: 2. Case studies in molecular design, *J. Am. Chem. Soc.* 119:4608–4617, 1997.

[21] M. Thormann and M. Pons, Massive docking of flexible ligands using environmental niches in parallelized genetic algorithms, *J. Comput. Chem.* 22:1971–1982, 2001.

[22] G. Jones, P. Willett, and R.C. Glen, Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation, *J. Mol. Biol.* 245:43–53, 1995.

[23] C.M. Oshiro, I.D. Kuntz, and J.S. Dixon, Flexible ligand docking using a genetic algorithm, *J. Computer-Aided Mol. Design* 9:113–130, 1995.

[24] P.G. Maillot, *Graphics Gems*, London: Academic Press, p. 498, 1996.

[25] T.J. Oldfield, A number of real-space torsion-angle refinement techniques for proteins, nucleic acids, ligands and solvent. *Acta Cryst.* D57:82–94, 2001.

[26] T.J. Oldfield, X-ligand: an application for the automated addition of flexible ligands into electron density, *Acta Cryst.* D57:696–705, 2001.

[27] G. Jones, P. Willett, R.C. Glen, A.R. Leach, and R. Taylor, Development and validation of a genetic algorithm for flexible docking, *J. Mol. Biol.* 267:727–748, 1997.

[28] V. Schnecke and L. Kuhn, Virtual screening with solvation and ligand-induced complementarity, *Persp. Drug Discov. Des.* 20:171–190, 2000.

[29] T.W. Whitfield, and J.E. Straub, Gravitational smoothing as a global optimization strategy, *J. Comput. Chem.* 23:1100–1103, 2002.

[30] U.H.E. Hansmann and L.T. Wille, Global optimization by energy landscape paving, *Phys. Rev. Lett.* 88:068105, 2002.

[31] W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery, *Numerical Recipes in Fortran*, Cambridge, UK: Cambridge University Press, 1992.

[32] H.J. Bohm, The development of a simple empirical scoring function to estimate the binding constant for a protein–ligand complex of known three-dimensional structure, *J. Computer-Aided Mol. Design* 8:243–256, 1994.

[33] M.D. Eldridge, C.W. Murray, T.R. Auton, G.V. Paolini, and R.P. Mee, Empirical scoring functions: I. the development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes, *J. Computer-Aided Mol. Design* 11:425–445, 1997.

[34] G.M. Verkhivker, D. Bouzida, D.K. Gehlhaar, P.A. Rejto, S. Arthurs, A.B. Colson, S.T. Freer, V. Larson, B.A. Luty, T. Marrone, and P.W. Rose, Binding energy landscapes of ligand–protein complexes and molecular docking: principles, methods, and validation experiments. In A.K. Ghose, and V.N. Viswanadhan, Eds., *Combinatorial Library Design and Evaluation: Principles, Software, Tools, and Applications in Drug Discovery*, New York: Marcel Dekker, pp. 157–195, 2001.

[35] M. Cecchini, P. Kolb, N. Majeux, and A. Caflisch, Automated docking of highly flexible ligands by genetic algorithms: a critical assessment, *J. Comput. Chem.* 25:415–422, 2004.

[36] B.R. Brooks, R.E. Bruccoleri, B.D. Olafson, D.J. States, S. Swaminathan, and M. Karplus, CHARMm: a program for macromolecular energy, minimization, and dynamics calculations, *J. Comput. Chem.* 4:187–217, 1983.

[37] W. Cornell, P. Cieplak, C. Bayly, I. Gould, K. Merz, Jr., D. Ferguson, D. Spellmeyer, T. Fox, J. Caldwell, and P. Kollman, A second generation force field for the simulation of proteins, nucleic acids, and organic molecules, *J. Am. Chem. Soc.* 117:5179–5197, 1995.

[38] M.L. Verdonk, J.C. Cole, P. Watson, V.J. Gillet, and P. Willett, SuperStar: improved knowledge-based interaction fields for protein binding sites, *J. Mol. Biol.* 307:841–859, 2001.

[39] I. Muegge, A knowledge-based scoring function for protein–ligand interactions: probing the reference state, *Persp. Drug Discov. Des.* 20:99–114, 2000.

[40] A.V. Ishchenko and E.I. Shakhnovich, Small Molecule Growth 2001 (SMoG2001): an improved knowledge-based scoring function for protein–ligand interactions, *J. Med. Chem.* 45:2770–2780, 2002.

[41] H. Gohlke, M. Hendlich, and G. Klebe, Knowledge-based scoring function to predict protein–ligand interactions, *J. Mol. Biol.* 295:337–356, 2000.

[42] P.S. Charifson, J.J. Corkery, M.A. Murcko, and W.P. Walters, Consensus scoring: a method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins, *J. Med. Chem.* 42:5100–5109, 1999.

[43] R.D. Clark, A. Strizhev, J.M. Leonard, J.F. Blake, and J.B. Matthew, Consensus scoring for ligand/protein interactions, *J. Mol. Graphics Modelling* 20:281–295, 2002.

[44] J. Warwicker and H.C. Watson, Calculation of the electric potential in the active site cleft due to α-helix dipoles, *J. Mol. Biol.* 157:671–679, 1982.

[45] M.K. Gilson and B.H. Honig, Energetics of charge-charge interactions in proteins, *Proteins: Structure, Function, and Genetics* 3:32–52, 1988.

[46] D. Bashford, and M. Karplus, pK_a's of ionizable groups in proteins: atomic detail from a continuum electrostatic model, *Biochem.* 29:10219–10225, 1990.

[47] M.E. Davis, J.D. Madura, B.A. Luty, and J.A. McCammon. Electrostatics and diffusion of molecules in solution: simulations with the University of Houston Brownian dynamics program, *Comput. Phys. Comm.* 62:187–197, 1991.

[48] W.C. Still, A. Tempczyk, R.C. Hawley, and T. Hendrickson, Semianalytical treatment of solvation for molecular mechanics and dynamics, *J. Am. Chem. Soc.* 112:6127–6129, 1990.

[49] M. Scarsi, J. Apostolakis, and A. Caflisch, Continuum electrostatic energies of macromolecules in aqueous solutions, *J. Phys. Chem.* A101:8098–8106, 1997.

[50] N. Budin, S. Ahmed, N. Majeux, and A. Caflisch. An evolutionary approach for structure-based design of natural and non-natural peptidic ligands, *Comb. Chem. High Throughput Screen.* 4:695–707, 2001.

[51] N. Budin, N. Majeux, C. Tenette-Souaille, and A. Caflisch, Structure-based ligand design by a build-up approach and genetic algorithm search in conformational space, *J. Comput. Chem.* 22:1956–1970, 2001.

[52] X. Zou, Y. Sun, and I.D. Kuntz, Inclusion of solvation in ligand binding free energy calculations using the generalized-Born model, *J. Am. Chem. Soc.* 121:8033–8043, 1999.

[53] N. Arora, and D. Bashford, Solvation energy density occlusion approximation for evaluation of desolvation penalties in biomolecular interactions, *Proteins: Structure, Function, and Genetics* 43:12–27, 2001.

[54] M. Rarey, B. Kramer, and T. Lengauer. The particle concept: placing discrete water molecules during protein–ligand docking predictions, *Proteins: Structure, Function, and Genetics* 34:17–28, 1999.

[55] J. Apostolakis, A. Plückthun, and A. Caflisch, Docking small ligands in flexible binding sites, *J. Comput. Chem.* 19:21–37, 1998.

[56] R.M.A. Knegtel, I.D. Kuntz, and C.M. Oshiro, Molecular docking to ensembles of protein structures, *J. Mol. Biol.* 266:424–440, 1997.

[57] R.M. Jackson, H.A. Gabb, and M.J.E. Sternberg, Rapid refinement of protein interfaces incorporating solvation: application to the docking problem, *J. Mol. Biol.* 276:265–285, 1998.

[58] P. Koehl and M. Delarue, Application of a self-consistent mean field theory to predict protein side-chains conformation and estimate their conformational entropy, *J. Mol. Biol.* 239:249-275, 1994.

[59] P. Koehl and M. Delarue, Mean-field minimization methods for biological macromolecules, *Curr. Opin. Struct. Biol.* 6:222–226, 1996.

[60] J.H. Lin, A.L. Perryman, J.R. Schames, and J.A. McCammon, Computational drug design accommodating receptor flexibility: the relaxed complex scheme, *J. Am. Chem. Soc.* 124:5632–5633, 2002.

[61] E. Yuriev and P.A. Ramsland, Mcg light chain dimer as a model system for ligand design: a docking study, *J. Mol. Recognit.* 15:331–340, 2002.

[62] J.D. Diller and C.L.M.J. Verlinde, A critical evaluation of several global optimization algorithms for the purpose of molecular docking, *J. Comput. Chem.* 20:1740–1751, 1999.

[63] A. Caflisch, P. Niederer, and M. Anliker, Monte Carlo docking of oligopeptides to proteins, *Proteins: Structure, Function, and Genetics* 13:223–230, 1992.

[64] M. Miller, S.K. Kearsley, D.J. Underwood, and M.D. Sheridan, FLOG — a system to select quasi-flexible ligands complementary to a receptor of known 3-dimensional structure, *J. Computer-Aided Mol. Design* 8:153–174, 1994.

[65] S. Makino and I.D. Kuntz, Automated flexible ligand docking method and its application for database search, *J. Comput. Chem.* 18:1812–1825, 1997.

[66] D.E. Goldberg, *Genetic Algorithms in Search Optimization and Machine Learning*, Reading, MA: Addison-Wesley, 1989.

[67] L. Davis, Ed., *Handbook of Genetic Algorithms*, New York: Van Nostrand Reinhold, 1991.

[68] D.K. Gehlhaar, G.M. Verkhivker, P.A. Rejto, C.J. Sherman, D.B. Fogel, L.J. Fogel, and S.T. Freer, Molecular recognition of the inhibitor AG-1343 by HIV-1 protease: conformationally flexible docking by evolutionary programming, *Chem. Biol.* 2:317–324, 1995.

[69] G.M. Verkhivker, P.A. Rejto, D.K. Gehlhaar, and S.T. Freer, Exploring the energy landscape of molecular recognition by a genetic algorithm: analysis of the requirements for robust docking of HIV-1 protease and FKBP-12 complexes, *Proteins: Structure, Function, and Genetics* 25:342–353, 1996.

[70] L. Schaffer and G.M. Verkhivker, Predicting structural effects in HIV-1 protease mutant complexes with flexible ligand docking and protein side-chain optimization, *Proteins: Structure, Function, and Genetics* 33:295–310, 1998.

[71] D. Hoffman, B. Kramer, T. Washio, T. Steinmetzer, M. Rarey, and T. Lengauer, Two-stage method for protein–ligand docking, *J. Med. Chem.* 42:4422–4433, 1999.

[72] J. Wang, P.A. Kollman, and I.D. Kuntz, Flexible ligand docking: a multistep strategy approach, *Proteins: Structure, Function, and Genetics* 36:1–19, 1999.

[73] M.L.P. Price, and W.L.J. Jorgensen, Analysis of binding affinities for celecoxib analogues with COX-1 and COX-2 from combined docking and Monte Carlo simulations and insight into the COX-2/COX-1 selectivity, *J. Am. Chem. Soc.* 122:9455–9466, 2000.

[74] W. Kabsch, A solution for the best rotation to relate two sets of vectors, *Acta Cryst.* A32:922–923, 1976.

[75] K. No, J. Grant, and H. Scheraga, Determination of net atomic charges using a modified partial equalization of orbital electronegativity method: 1. Application to neutral molecules as models for polypeptides, *J. Phys. Chem.* 94:4732–4739, 1990.

[76] K. No, J. Grant, M. Jhon, and H. Scheraga. Determination of net atomic charges using a modified partial equalization of orbital electronegativity method: 2. Application to ionic and aromatic molecules as models for polypeptides, *J. Phys. Chem.* 94:4740–4746, 1990.

[77] K. Hilpert, J. Ackermann, D.W. Banner, A. Gast, K. Gubernator, P. Hadvary, L. Labler, K. Müller, G. Schmid, T. Tschopp, and H. van de Waterbeemd, Design and synthesis of potent and highly selective thrombin inhibitors, *J. Med. Chem.* 37:3889–3901, 1994.

[78] S. Piana, D. Sebastiani, P. Carloni, and M. Parrinello, *Ab initio* molecular dynamics-based assignment of the protonation state of pepstatin A/HIV-1 protease cleavage site, *J. Am. Chem. Soc.* 123:8730–8737, 2001.

[79] M.L. Raymer, P.C. Sanschagrin, W.F. Punch III, S. Venkataraman, E.D. Goodman, and L.A. Kuhn, Predicting conserved water and water-mediated ligand interactions in proteins using a k-nearest-neighbor genetic algorithm, *J. Mol. Biol.* 265:445–464, 1997.

Chapter 3

Automated Docking of Highly Flexible Ligands by Genetic Algorithms: A Critical Assessment

(Journal of Computational Chemistry 25, pp 412-422, 2004)

Automated Docking of Highly Flexible Ligands by Genetic Algorithms: A Critical Assessment

MARCO CECCHINI, PETER KOLB, NICOLAS MAJEUX, AMEDEO CAFLISCH
Department of Biochemistry, University of Zürich, Winterthurerstrasse 190, CH-8057 Zürich, Switzerland

Received 30 May 2003; Accepted 6 August 2003

Abstract: An improved version of the fragment-based flexible ligand docking approach SEED–FFLD is tested on inhibitors of human immunodeficiency virus type 1 protease, human α-thrombin and the estrogen receptor β. The docking results indicate that it is possible to correctly reproduce the binding mode of inhibitors with more than ten rotatable bonds if the strain in their covalent geometry upon binding is not large. A high degree of convergence towards a unique binding mode in multiple runs of the genetic algorithm is proposed as a necessary condition for successful docking.

 J Comput Chem 25: 412–422, 2004

Key words: docking; genetic algorithm; SEED; FFLD; ligand flexibility; HIV-1 protease

Introduction

Computer-aided approaches for docking small molecules into proteins of known structure are useful tools for drug design.[1–4] The importance of automatic docking procedures keeps growing because of the ever increasing number of 3D structures of pharmacologically relevant enzymes and receptors. Further, combinatorial and parallel synthesis techniques have generated a significant number of libraries of compounds with good pharmacological properties[5] and in certain cases tailored for specific targets.[6,7] Automatic approaches are available for docking flexible ligands of up to about 10 rotatable bonds into rigid[8–10] and partially flexible targets,[11–17] and several review articles have been published.[2,18–20] Ligands with a larger amount of rotatable bonds are much more difficult to dock,[13,14] even using a rigid protein.[21,22]

In this article we present the improvements with respect to the original version of SEED–FFLD[10,23] and evaluate the new version on difficult test cases. The SEED–FFLD approach is based on the decomposition of ligands into mainly rigid fragments. First, the most favorable fragment positions and orientations in the receptor binding site are determined by the program SEED according to an accurate binding energy that includes electrostatic solvation effects.[24] The optimal binding modes of the fragments are then used in FFLD as binding site descriptors to guide the placement of ligand conformations generated by a genetic algorithm (GA). The SEED–FFLD approach was tested by docking known nanomolar inhibitors with about 10 rotatable bonds in the active site of the uncomplexed and complexed conformations of thrombin and the human immunodeficiency virus type 1 protease (HIV-1 PR).[10]

The present work was motivated by two main questions: Is it possible to dock ligands with more than 10 rotatable bonds into HIV-1 PR? Are the predicted binding modes affected by the protein conformation (choice of the crystal structure)? The docking results show that redocking is almost always successful whereas cross-docking is problematic mainly because of strain in the covalent geometry of the ligand. Large and flexible ligands might be of limited relevance in the context of drug design. Yet, we think that testing docking programs in cases where the conformational space is very large is useful to find weaknesses and suggest improvements that will be in any case beneficial also for smaller and/or more rigid ligands.

Materials and Methods

The docking approach is a three-step strategy based on the decomposition of a flexible ligand into rigid fragments. First, the program SEED is used to dock the fragments into the binding site of the receptor.[23,25] Second, the ligand is docked by a genetic algorithm (FFLD) that uses a fast scoring function.[10] The genetic algorithm perturbations affect only the conformation of the ligand; its placement in the binding site is determined by the SEED anchors and a least-square fitting method.[26] In this way the position and orientation of the ligand in the binding site are determined by the best binding modes of its fragments previously docked using an accurate energy function with electrostatic solvation.[24] The scoring

Correspondence to: A. Caflisch; e-mail: caflisch@bioc.unizh.ch

function used in FFLD is based on van der Waals and hydrogen bond terms and does not explicitly include solvation for efficiency reasons. Solvation effects are implicitly accounted for as the binding modes of the fragments are determined with continuum electrostatics.

Third, the FFLD results are postprocessed by CHARMM minimization. The ligand decomposition into fragments, the choice of functional groups (preferable large aromatic rings) as anchors, the identification of rotatable bonds, and the definition of the binding site have recently been automated (P. Kolb et al., unpublished results). The other modifications of the SEED–FFLD approach with respect to the original procedure[10] are listed in the next subsections.

SEED

Fragment Docking

The SEED input parameters used for this application are identical to those in Table I of the original SEED article[23] except for the following three: (1) The interior dielectric constant is set to 2.0 to partially account for the electronic polarizability and dipolar reorientation effects of the solute. (2) The number of apolar points on the receptor is increased from 100 to 300 because of the very large buried binding site of HIV-1 PR. For human α-thrombin and the estrogen receptor β 300 and 150 apolar points were used, respectively. (3) To discard polar and apolar receptor vectors that point outside of the binding site, a selection using an angle criterion is performed. Initially, the minimal and maximal distances between the end points of the vectors and a set of points in the binding site (e.g., the positions of the heavy atoms of the ligand) are evaluated. A vector is discarded if the angle it spans with the closest point is larger than a cutoff. This selection uses a permissive cutoff of 100° for vectors close to the binding site points and a stricter one (70°) for distant vectors. Using the atoms of a ligand from a known complex to define the binding site does not introduce a bias in cross-docking and corresponds to the situation in an advanced drug design program, where one or more crystal structures of protein–ligand complexes have already been solved.

Postprocessing of the Optimal Binding Modes of the Fragments

The fragment binding modes are sorted by binding energy and clustered in SEED according to their position and orientation in the binding site using a conservative criterion based on distances between similar atom types.[23,27,28] Cluster representatives are subsequently grouped according to the coordinates of the geometric centers using a threshold of 1 Å. The geometric centers of the first five cluster representatives are removed from the clustering procedure described in the following and directly used for docking. For each cluster an average geometric center ($\mathbf{r}_{\mathrm{AGC}}$) is calculated with the following procedure. First, all the positions with a binding energy greater than 3 kcal/mol with respect to the cluster representative are discarded. Then, the $\mathbf{r}_{\mathrm{AGC}}$ of a given cluster is evaluated as an energy-weighted mean

$$\mathbf{r}_{\mathrm{AGC}} = \sum_{i=1}^{N} \omega_i \mathbf{r}_i,$$

$$\omega_i = \begin{cases} E_i \Big/ \sum_i E_i & \text{if } E_{\max} < 0 \\ (E_i - (E_{\min} + E_{\max})) \Big/ \sum_i (E_i - (E_{\min} + E_{\max})) & \text{if } E_{\min} > 0 \end{cases} \quad (1)$$

where $E_{\min}$ and $E_{\max}$ are the minimum and maximum energy within a cluster, respectively, and the sum runs over the N geometric centers $\mathbf{r}_i$ of the fragments in the cluster. In the sporadic case of $E_{\min} > 0$, by subtracting $(E_{\min} + E_{\max})$ from the fragment binding energies E_i it is possible to give more weight to positions with small absolute energy values. For $E_{\min} \leq 0 \leq E_{\max}$, average centers for the subsets of favorable ($E_i < 0$) and unfavorable ($E_i > 0$) binding modes in a cluster ($\mathbf{r}_-$ and $\mathbf{r}_+$, respectively) are computed by eq. (1) and the $\mathbf{r}_{\mathrm{AGC}}$ is evaluated as

$$\mathbf{r}_{\mathrm{AGC}} = 0.8\mathbf{r}_- + 0.2\mathbf{r}_+ \quad (2)$$

where the multiplicative factors 0.8 and 0.2 are somewhat arbitrary and were not optimized. The first 15 $\mathbf{r}_{\mathrm{AGC}}$ are kept for FFLD. Therefore, the binding site maps used for docking are defined by 20 points, 5 corresponding to geometric centers of fragments optimally docked by SEED and 15 average geometric centers. In this way, the final list of geometric centers used for docking is a compromise between the accuracy of the SEED binding energy and the diversity derived from the clustering procedure. The postprocessing of the optimal binding modes of the fragments leads to more heterogeneous binding site maps than using only the most favorable ones. To store this information for efficient placement of the ligand in the binding site three hash tables are generated as described in ref. 10.

FFLD

The following subsections contain the improvements with respect to the original version of FFLD.[10] The FFLD scoring function consists of an intraligand van der Waals energy, a ligand–protein soft core van der Waals, and an intermolecular polar energy term.[10] The two latter terms were modified as described below.

Intermolecular Soft Core van der Waals

A soft core van der Waals is used to generate a smooth energy landscape by reducing the frustration originating from the steep repulsive part.[29] In the present work, the linearization of the Lennard–Jones potential used to smooth the energy surface was slightly modified with respect to ref. 10. Originally, upon ligand placement, for every atom located in the binding site (including a 1-Å layer below the protein surface) a van der Waals energy with the closest protein atom was evaluated at first. The interaction energy was linearized if its value was higher than a cutoff; otherwise, a grid-based interpolated interaction energy including all

contributions of the receptor atoms was considered. In the original procedure small atomic interpenetrations with the protein surface were overpenalized without taking into account the favorable contributions of the neighboring atoms. In the version used in the present study, the van der Waals interaction with the closest protein atom is compared to the grid-based interpolated interaction energy. If the latter is more favorable, the attractive contributions of the receptor atoms dominate the interaction and the grid-based energy is more appropriate. Otherwise, the atomic interpenetration is significant and the repulsive contribution dominates the interaction. In this case, the linearized interaction with the closest protein atom is used.

Polar Interactions

The polar term approximates electrostatic interactions between ligand and protein:

$$E_{\text{polar}}^{\text{inter}} = \sum_{i=1}^{N_{\text{HB}}} E_i^{\text{HB}} + \sum_{i=1}^{N_{\text{UP}}} E_i^{\text{UP}} \tag{3}$$

where N_{HB} and N_{UP} are the number of hydrogen bonds (HBs) and unfavorable polar contacts (UPs), respectively. A significant improvement with respect to the original version[10] is the replacement of step functions, which allow different binding modes with the same energy value with smooth functions. Smooth functions allow the optimization of the hydrogen bonding pattern, avoiding discontinuities on the energy surface. A dependence on the distance and the angle in the polar term of the scoring function was introduced to reduce the noise arising from the energy degeneration and improve the convergence of the docking runs. The criteria used for the definition of unfavorable polar contacts and hydrogen bonds and the smooth functions implemented are described below.

Unfavorable Polar Contacts. An interaction between two H-bond donors or two H-bond acceptors is an unfavorable polar contact whose energy is a function of the distance between the interacting heavy atoms. A sigmoidal function is used

$$E^{\text{UP}}(r) = \begin{cases} E_{\text{bad}} & r < r_{\text{on}} \\ E_{\text{bad}}(1 + e^{\beta(r-\alpha)})^{-1} & r_{\text{on}} \le r \le r_{\text{off}} \\ 0 & r > r_{\text{off}} \end{cases} \tag{4}$$

where r is the distance between the heavy atoms, E_{bad} is the maximal penalty for an unfavorable polar contact, α is the interatomic distance corresponding to the inflection point of the function, and β is related to the steepness of the sigmoidal. The value of $E_{\text{bad}} = 3.0$ kcal/mol is the same as used previously.[10] The values of r_{on} and r_{off} depend on the choice of α and β, which were fixed at 2.8 Å and 15.5 Å^{-1}, respectively. The value of r_{on} is defined as the point where the sigmoidal reaches a value of 0.99 E_{bad}. For symmetry reasons, r_{off} is fixed by the choice of r_{on}. Hence, $r_{\text{off,on}} = \alpha \pm 4.65/\beta$.

Hydrogen Bonds. The hydrogen bonding model of the MAB force field[30]

$$E^{\text{HB}}(r, \theta) = W_{\text{H}} B_R(r) \Theta_D(\theta) \tag{5}$$

was implemented, where r is the distance between the donor (D) and the acceptor (A) atoms, θ is the angle at the H atom (D—H $\cdots$ A), and W_{H} is an atom-type dependent parameter that defines the strength of the bond. The energy is determined by the parameter W_{H}, while the distance dependence and the directionality of the hydrogen bonds follow bathtub-shaped functions:

$$B_R(r) = -\left[1 - \left(\frac{r - r_{\text{eq}}}{w_r}\right)^n\right]^m \tag{6}$$

$$\Theta_D(\theta) = \left[1 - \left(\frac{\theta - \theta_{\text{eq}}}{w_\theta}\right)^n\right]^m \tag{7}$$

Equations (6) and (7) apply whenever the expression in the outer brackets is positive; otherwise, B_R and Θ_D vanish. The exponents n and m determine the curvature of the bathtub-shaped functions and were chosen as $n = 2$ and $m = 4$. Compared to ref. 30, the implemented model, while preserving an all-atom description, does not take into account the angle at the acceptor atom in the hydrogen bond energy evaluation. Although this could lead to a loss in accuracy, it avoids the need for additional parameters for describing the valence state of the atoms involved in the hydrogen bond. The following parameters were used: the well depth W_{H} and the optimal distances r_{eq} were chosen according to the atom types of the donor and acceptor atoms involved;[23,30] θ_{eq} was set to π because of the linearity of optimal hydrogen bonds; w_r was 5.0 Å when $r \le r_{\text{eq}}$ or 1.25 Å when $r > r_{\text{eq}}$ and w_θ was 1.5 radian to mimic the original stepwise functions.

Local Search

Following a comparative study of several search engines for AutoDock,[31] a hybrid search procedure was implemented in FFLD. The hybrid search combines a global optimization procedure based on a genetic algorithm to overcome energy barriers with a local minimization algorithm to explore regions within energy basins. Local optimization has been shown to dramatically improve the success rate of the genetic algorithm search without any loss in efficiency.[31] The use of a local optimizer increases the fitness of the individuals and accelerates convergence. In the hybrid search, a loop over generations is performed until the maximum number of generations or the maximum number of energy evaluations is reached. In a generation, five different stages follow one another: evolution, mapping, fitness evaluation, local search, and similarity test. At the beginning of each cycle, the *old population* is evolved by means of the genetic operators, one-point crossing over and mutation, and a *new population* is generated. The new genetic material is decoded and the binding energy is evaluated. For the best 10% of the individuals, the local search is performed to improve the ligand fitness. Finally, the individuals of the *old population* are replaced by new chromosomes, taking into account both the energy difference and the conformational similarity.[32] The latter dramatically improves the efficiency of the hybrid search because the local search can easily cause the system to get trapped in local minima. In fact, to avoid premature con-

vergence it is important to keep structural diversity during the evolution. As a local optimizer, the Solis and Wets algorithm[33] was used with the following parameters: The maximum number of iterations per search was set to 300; the local search stepsize (ρ) was 0.1 radian; the maximum number of consecutive successes before increasing ρ was 5, while the maximum number of consecutive failures before decreasing ρ was 3; the lower bound on ρ, i.e., the termination criterion for the local search, was 0.01.

Postprocessing by CHARMM Minimization

For every docking experiment, 10 genetic algorithm runs were performed. For each run, the binding mode with the lowest FFLD energy was postprocessed by CHARMM minimization[34] using the CHARMm22 force field (Accelrys, Inc.). The structure of the protein, including the critical bridging water, was held fixed. A distance-dependent dielectric function [$\varepsilon(r) = r$] was used and the conjugate gradient minimization was stopped when the root mean square of the energy gradient reached a value of 0.01 kcal mol^{-1} $Å^{-1}$. The CHARMM-minimized ligand conformations were sorted according to their binding energy and the lowest-energy solution was compared to the reference ligand (see Preparation of the Ligands section below). The heavy atom root mean square deviation (RMSD) from the reference was determined as a quantitative measure of the docking reliability. No least-square superposition is used in calculating the RMSD because the rigid protein establishes a fixed frame of reference.

System Setup

Test Cases

To evaluate the performance of the improved version of the SEED–FFLD procedure, five HIV-1 PR protein–ligand complexes were considered. Inhibitors of HIV-1 PR are peptidomimetic molecules with a dozen or more rotatable bonds and, as such, they present a challenging target for automated docking techniques. Moreover, a large number of crystallographic structures of HIV-1 PR protein–ligand complexes are available from the Protein Data Bank (PDB) database.[35] The crystal structures used in this study were 1hvr, 1hbv, 1htg, 1hvs, and 5hvp.[36–40] The corresponding set of ligands is characterized by a wide range of torsional degrees of freedom (between 10 and 22 rotatable bonds) and contains a certain amount of diversity due to the different functional groups (Fig. 1). The ligand 1hvr has the lowest number of rotatable bonds (10) and is the only nonpeptidic inhibitor of the set. It includes aromatic fragments suitable as anchors and contains a large central scaffold, the cyclic urea, which significantly reduces the complexity of its conformational space by decoupling most of the torsional degrees of freedom. Moreover, the cyclic urea itself strongly interacts with the protein, forming on one side hydrogen bonds with the flaps and on the other side hydrogen bonds with the aspartyl dyad. Therefore, the ligand 1hvr is expected to be the simplest test case of the set. The ligands 1hbv and 1htg present a larger amount of flexibility (15 and 17 rotatable bonds, respectively) with respect to 1hvr and therefore represent a test system having an intermediate level of difficulty. Eventually, the ligands

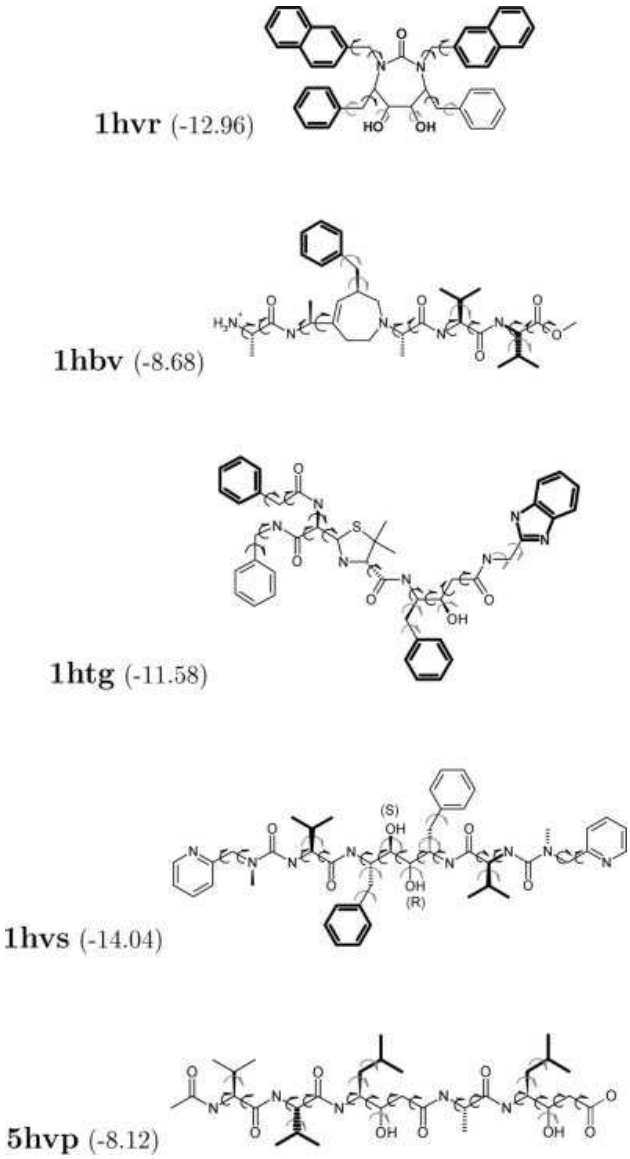


Figure 1. HIV-1 PR inhibitors used in this work. Coordinates for each complex were obtained from the PDB[35] using the accession codes given in bold. Bonds that were treated as flexible are marked by curly arrows; gray and black arrows indicate side-chains[14] and main-chain, respectively. Fragments used in SEED are in bold. Experimental free energies of binding (kcal/mol) are given in parentheses; they were obtained from the primary reference for each crystallographic structure.[36–40]

1hvs and 5hvp (21 and 22 rotatable bonds, respectively) provide very difficult test cases because of the large conformational space.

In addition, the *N*α-((2-naphthylsulfonyl)glycyl)-DL-*p*-amidinophenylalanyl-piperidine (NAPAP)/human α-thrombin complex[41] and the complex between lig15 and the estrogen receptor β[42] were investigated to measure the robustness of the docking method. The former is a cross-docking experiment in the uncomplexed structure of the

human α-thrombin (1hgt). The latter is a redocking simulation in the "native" conformation of the estrogen receptor β (1nde). NAPAP is a peptidomimetic molecule with eight rotatable bonds. It includes two large hydrophobic fragments, naphthalene and piperidine, and the positively charged benzamidine. Lig15 is a 1, 3, 5-triazine-based molecule with 11 rotatable bonds. The ligand presents a trisubstituted planar central scaffold (the triazine) with flexible linkers to hydrophobic substituents.

Preparation of the HIV-1 PR Structures

The five crystal structures were downloaded from the PDB database.[35] The ligand and all water molecules but one were removed. The water bridging the two flaps was retained except for the docking runs with either the 1hvr protease structure or the 1hvr ligand (cyclic urea inhibitor). The side-chains of lysine and arginine residues were protonated, as well as the side-chains of histidine residues because the optimal pH of HIV-1 PR is around 5 and they are exposed to the solvent. The carboxylate groups of aspartic and glutamic acid were ionized. Particular attention was addressed to the ionization state of the cleavage site, which contains the aspartyl dyad (Asp25/Asp25′). At optimal pH for enzymatic activity (~5–6), the aspartyl dyad is most probably monoprotonated in the uncomplexed enzyme. Besides the pH, different inhibitors can have an effect on the ionization state of the active site because they can stabilize the neutral dyad or the negatively or dinegatively charged forms. According to Piana et al.,[43] the monoprotonated state is accompanied by the presence of two strong hydrogen bonds between the aspartyl dyad and H-bond donors belonging either to the inhibitor or to an ordered water molecule. Hence, a monoprotonated state was considered for the proteases 1hvr and 1hvs, while a diprotonated state was chosen for the others. Hydrogen atoms were added to the structures and minimized with the program CHARMM[34] and the CHARMm22 force field (Accelrys, Inc.). Partial charges were assigned using the MPEOE method.[44,45] Finally, the binding site was defined as the smallest zone that encompasses the HIV-1 PR residues with more than 50% of the atoms within a distance of 5 Å from any atom of the inhibitor in the X-ray structure of the complex.

Preparation of Human α-Thrombin and the Estrogen Receptor β

After downloading the crystal structures from the PDB database,[35] human α-thrombin (1hgt) and the estrogen receptor β (1nde) were prepared following a procedure similar to the one described above for HIV-1 PR.

Human α-thrombin is a trypsin-like serine protease that fulfills a central role in both hemostasis and thrombosis.[46] Several inhibitors are known to bind to the nonprime region of the active site, i.e., pockets S3–S1. The S3 and S2 pockets are hydrophobic. S3 is occupied by an aromatic ring in most of the known inhibitors and by the Phe side-chain in the natural substrate. The S2 pocket is usually filled by aromatic or aliphatic side-chains. The S1 pocket is a cylindrical cavity with an Asp at the bottom. It is usually filled by a positively charged side-chain (Arg, Lys, benzamidine, etc.) involved in a salt bridge with the Asp. During the preparation of the protein, particular attention was addressed to the protonation state of the active site residues (His, Asp, Ser). In particular, the catalytic His was protonated at the δ nitrogen. Its monoprotonated state is fully compatible with its partially buried position in the binding site and allows the formation of two intramolecular hydrogen bonds between the residues of the catalytic tryad.

The estrogen receptor β is a ligand-activated transcription factor that plays a key role in the modulation of gene expression. It binds a wide range of steroidal and nonsteroidal ligands with moderate to high affinity, with the minimal requirement of at least one paramonosubstituted phenol as the basic pharmacophore.[42] The estrogen receptor β binding site is almost buried and predominantly hydrophobic. Nevertheless, the paramonosubstituted phenol increases the ligand binding affinity because its hydroxyl group fits optimally into the gap between Arg394 and Glu353, accepting a hydrogen bond from the guanidinium and donating a hydrogen bond to the carboxy group.[42] In the structure downloaded from the PDB database,[35] the coordinates of 24 residues were missing due to poor electron density. These incomplete regions were far away from the binding site and were neglected in the docking experiments.

Preparation of the Ligands

The initial coordinates were downloaded from the PDB database.[35] Formal charges were assigned by ionizing carboxylic-acid groups and protonating amino groups. Hydrogen atoms were added and partial charges were assigned (see previous subsection). The molecular structure was minimized in two different ways with the program CHARMM[34] and the CHARMm22 force field (Accelrys, Inc.).

The first minimization was carried out in the binding site with the protein fixed. For redocking, the ligand structure optimized within its own protein conformation is used as reference for the calculation of the RMSD values. For cross-docking, protein coordinates were first superimposed by fitting the C_α atoms. The ligand conformation then optimized in the "non native" protein is used as the RMSD reference structure. In both cases, the reference structures used for evaluating docking results differ from the experimental conformation of the bound ligand. Nevertheless, the comparison between reference and predicted conformations is more consistent in this way because both structures correspond to minima of the same energy function (CHARMm22, Accelrys, Inc.). For redocking, the RMSD between X-ray and reference structures was 0.4, 0.6, 0.7, 1.0, 1.1, and 1.0 Å for 1hvs, 1htg, 1hvr, 1hbv, 5hvp, and 1nde, respectively. The average RMSD between X-ray and reference structures for cross-docking was 1.3 ± 0.4 Å, with a maximum value of 2.4 Å.

The second ligand minimization was performed outside of the receptor to remove any bias originating from the PDB structure of the protein–ligand complex [see type (3) docking experiments below]. Both minimizations were carried out using the same protocol as described before for the postprocessing.

Results and Discussion

The results on HIV-1 PR are discussed first while the human α-thrombin and estrogen receptor β results are at the end of this

section. Each of the five HIV-1 PR ligands was docked to each of the 5 protein structures, yielding a matrix of 25 docking experiments. The complexes along the matrix diagonal correspond to redocking experiments and are expected to more easily match the crystallographic structures because any "induced fit" due to the inhibitor is already present in the protein conformation. To study the effect of both ligand flexibility and geometry (covalent bond distances and bond angles in the ligand input structure) on the docking results three kinds of docking experiments were performed:

1. ***Biased geometry and partial flexibility.*** The covalent geometry of the ligand was minimized with CHARMM in the binding site of the receptor before running FFLD. The docking experiments were performed with flexible ligand side-chains and rigid main-chain (same rotatable bonds as in ref. 14; gray arrows in Fig. 1).
2. ***Biased geometry and full flexibility.*** The covalent geometry of the ligand was the same as in (1) but the experiments were carried out allowing flexibility to all rotatable bonds (gray and black arrows in Fig. 1).
3. ***Unbiased geometry and full flexibility.*** The ligand structure was minimized with CHARMM outside of the receptor to remove any geometric bias. The ligand flexibility was the same as in (2).

Conformations docked within 2.4 Å heavy-atom RMSD from the reference structure are considered successes.

Biased Geometry and Partial Flexibility

In docking experiments (1), the SEED–FFLD procedure was able to correctly dock the five ligands in each of the five protein structures (Fig. 2, top). Upon CHARMM minimization, the lowest-energy conformation for each experiment reproduced the crystallographic binding mode very well with a maximal RMSD of 1.1 Å with respect to the reference structure. Moreover, all of the 10 docking runs gave the same binding mode in 23 of 25 cases. Lack of full convergence for inhibitors 1hbv and 1htg in the 1hvr HIV-1 PR conformation was probably due to the lack of the water bridging the flaps in the 1hvr structure of the protease. This structural water plays an important role in the molecular recognition process, acting as anchor in the active site. Note that the CHARMm energy of the docking solution can be up to 16 kcal/mol (1htg in 5hvp) more favorable than the energy of the reference structure despite RMSD values smaller than 1.2 Å (Fig. 2, top). This is mainly due to electrostatic contributions coming from a better placement of the hydroxyl hydrogen interacting with the catalytic dyad (hydrogens are not considered in the RMSD evaluation). Albeit successful, the docking experiments of type (1) assume the knowledge of the backbone conformation of the bound ligand and are therefore only marginally useful (e.g., to dock a library of compounds with the same backbone).

Biased Geometry and Full Flexibility

In docking simulations of fully flexible ligands, such as experiments (2) and (3), the conformational space of the ligand is much larger than in (1). Nevertheless, in experiments (2) redocking and cross-docking of inhibitors 1hvr and 1hvs gave good results while cross-docking of 1hbv, 1htg, and 5hvp was only partially successful.

Cross-docking with the ligand 5hvp and with the protease 1hvr proved difficult for the SEED–FFLD procedure. In the first case, the large conformational space of the ligand (22 rotatable bonds) and the lack of fragments suitable as anchors were crucial for the poor performance of the simulations. Moreover, the wrong binding modes have a CHARMm energy more favorable than the reference structure (Fig. 2, right) and this points to limitations of the energy model. As an example, the misdocked conformation of the ligand 5hvp in the protease 1hvr (RMSD of 3.8 Å and $E_{\mathrm{pred}}^{\mathrm{CHARMm}} - E_{\mathrm{ref}}^{\mathrm{CHARMm}} = -18.9$ kcal/mol) is completely bent in the middle, folds back onto itself, and occupies only half of the binding site "cylinder." Although both inter- and intramolecular van der Waals interactions are optimized by this binding mode, the negatively charged terminal carboxy group is buried. This is a clear indication that solvation is needed for the final ranking of the solutions, especially for very flexible ligands without anchors.

The lack of the structural water in the protease 1hvr precluded the reproduction of the experimental results. Although successful, most of the redocking and cross-docking experiments did not converge in all cases (Fig. 2, middle). This is due to the large space accessible to ligands and in particular to the presence of flexibility in the main-chain. In fact, only the convergence for the ligand 1hvr, which has a rigid central ring rather than a flexible main-chain, was not affected by the increased number of rotatable bonds.

Unbiased Geometry and Full Flexibility

As for the docking experiments of type (1), experiments (2) require the knowledge of the bound complex, at least for determining the geometry of the input ligand structure. Therefore, to reproduce the typical high-throughput docking conditions completely unbiased docking simulations, experiments (3), were performed. Here, redocking was successful in three of five cases (1hvr, 1htg, and 1hvs), while cross-docking gave good results only for the highly flexible ligand 1hvs (Fig. 2, bottom). Experiments of type (3) are much more prone to fail with respect to experiments of type (2), even though both deal with the same amount of ligand dihedral flexibility. With respect to this point, the docking simulations of the ligands 1hvr and 1hbv are in particular significant. For the 1hvr inhibitor, while experiments (2) always converged to the reference structure experiments (3) completely misdocked the ligand in the majority of the cross-docking simulations. In the case of ligand 1hbv, while partially successful in experiments (2) and (3) did not reproduce the experimentally determined binding mode.

The explanation for this clear worsening in the performance of docking calculations is to be searched for in the covalent geometry of the ligand structures used as input in experiments (2) and (3). Therefore, for each docking simulation the biased and unbiased input structures were superimposed and compared. In most cases, the biased structure, i.e., the geometry of the bound ligand, was far away from the energy minimum and the CHARMM minimization performed outside of the receptor yielded geometries not compatible with the binding site (Fig. 3). In the ligand 1hvr, the covalent angle at the methylene carbons linking the naphthalenes to the

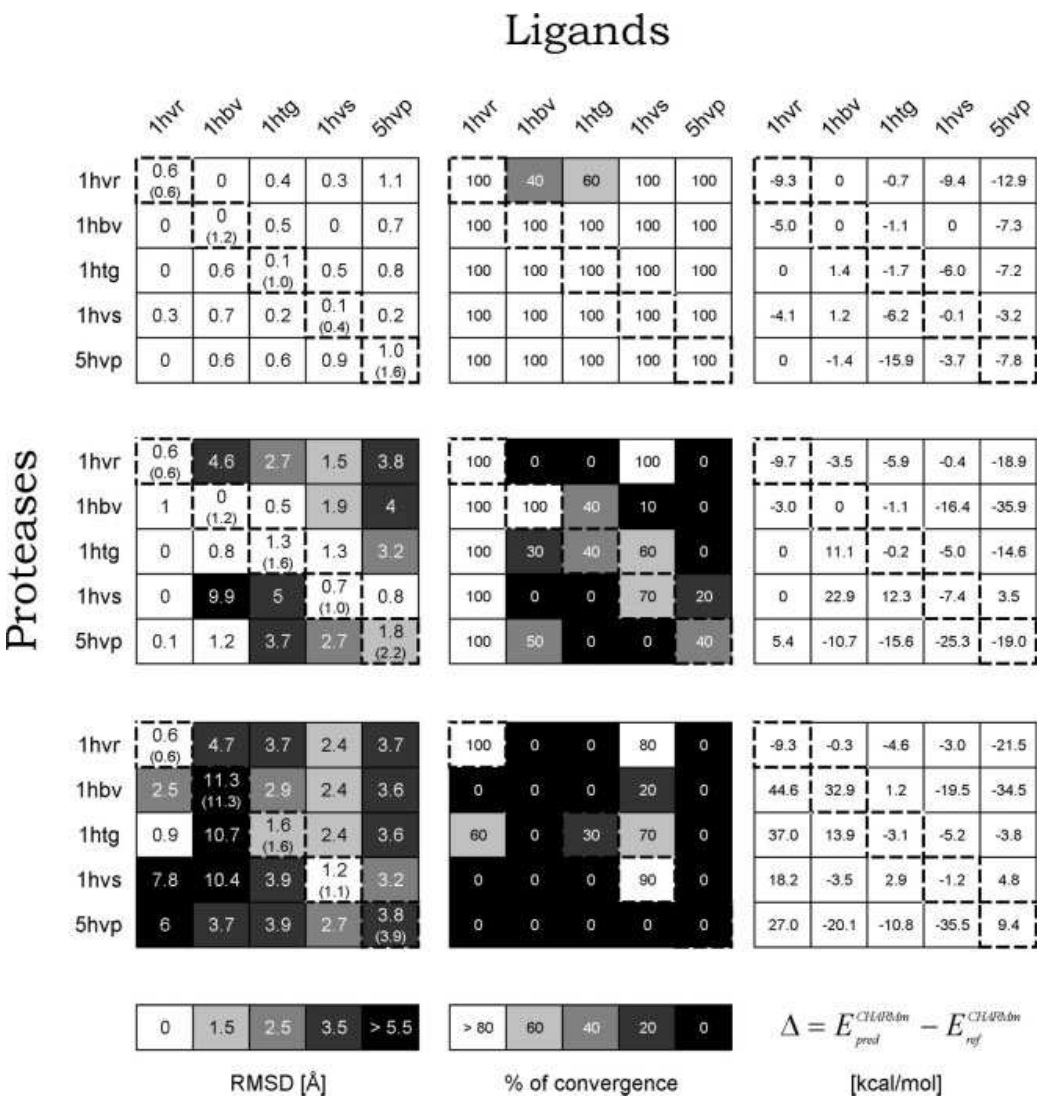


Figure 2. Results of the docking simulations. Redocking (boxes along the diagonals) and cross-docking experiments of types (1), (2), and (3) are presented from top to bottom, respectively. In the left column, heavy-atom RMSD values between the docked conformation of most favorable energy and the reference structure are given (Å). For redocking experiments, RMSD values to the unminimized experimentally determined ligand positions are reported in parentheses. In the middle column, the convergence of docking experiments are given in terms of percent ratio of successes. Docking predictions within 2.4 Å RMSD of the reference structure were considered successes. In the right column, the CHARMm energy difference between the docking solution and the reference structure is reported. Large negative values indicate limitations in the scoring function whereas positive values may point to uncomplete sampling.

cyclic urea is stretched to about 120° in the bound conformation to optimize the van der Waals interactions between the naphthalenes and the protein. In the minimization outside of the receptor, the covalent angle at the methylene carbons relaxes to a value of about 113°.

Analogous considerations can be made for the ligand 1hbv (Fig. 3). Here, the geometry of the nitrogen atom belonging to the cyclic scaffold is stretched to about 120° in the bound conformation. Upon minimization outside of the receptor, the geometry at the nitrogen atom relaxes and moves to a more pronounced tetrahedral geometry. In both cases, the CHARMM minimization outside of the receptor significantly modifies the covalent geometry of the ligands so that the SEED–FFLD approach cannot reproduce the experimental binding mode only by dihedral modifications of the input structure.

For redocking 1hvr and cross-docking of 1hvr into 1htg, this problem was overcome by CHARMM postprocessing. The lowest-energy conformation found by the SEED–FFLD procedure was

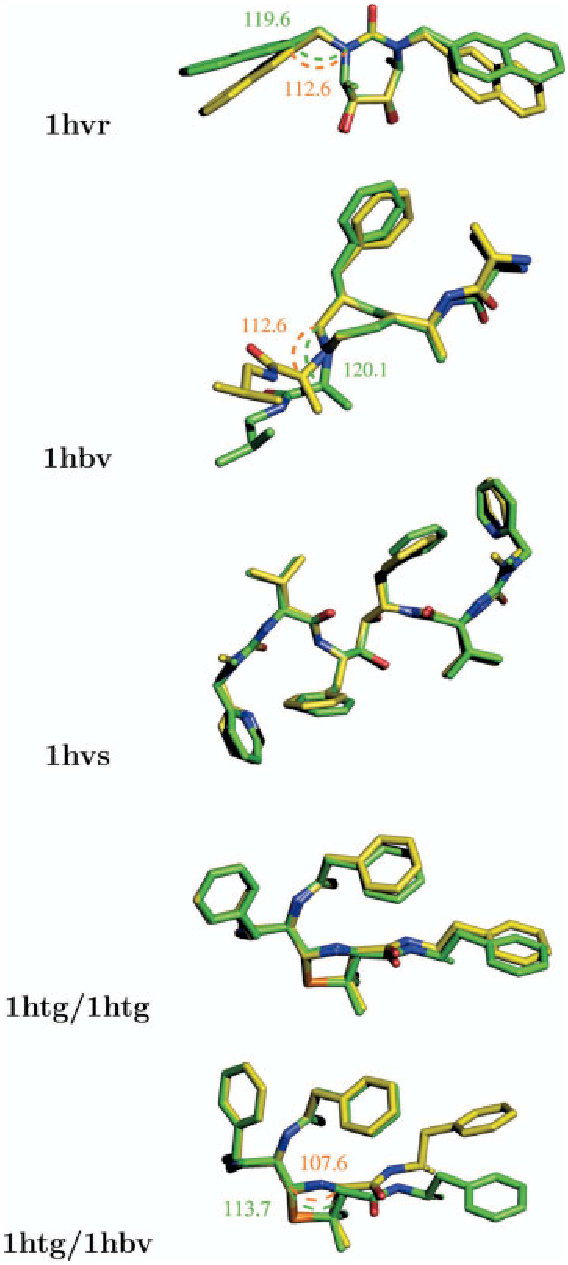


Figure 3. Comparison between ligand structures with biased geometry (green carbons) and unbiased geometry (yellow carbons) used as input for redocking 1hvr, 1hbv, 1hvs, and 1htg and for cross-docking 1htg with the protease 1hbv (from top to bottom, respectively). Significant deviations in the covalent angles are marked by broken arcs. (The pictures of the ligands were drawn using the program PyMOL.[48])

close to the X-ray structure so that during CHARMM postprocessing the experimental binding mode was reproduced. However, it has to be mentioned that this was a fortuitous event. In fact, cross-docking of the 1hvr inhibitor failed in three of four experiments (Fig. 2, bottom). It is important to note that the CHARMm energy of the docking solution is much poorer than the one of the reference structure not because the sampling in dihedral space was incomplete. Rather, the reason is that the optimal covalent geometry (obtained by minimization outside of the binding site) prevents the FFLD docking from reaching the basin of the CHARMm energy that contains the X-ray structure. Hence, the docking procedure has no chance to succeed, even after CHARMM postprocessing. This happened also with the ligand 1hbv, where even redocking was not successful. Here, the apparently small deformation of the central ring (Fig. 3) played a crucial role in docking and was the main reason for the observed failures.

The ligand 1hvs has similar biased and unbiased input structures (Fig. 3), i.e., there is not any significant deformation in the bound conformation. Hence, the unbiased geometry cannot prevent SEED–FFLD from finding a solution close to the correct one and the performance of experiments (2) and (3) are comparable. However, the biased and unbiased geometry structures are not identical and the ligand poses predicted by experiments (3) are slightly worse than those in (2) (Fig. 2).

Another interesting case is the redocking of the ligand 1htg and its cross-docking with the protease 1hbv. As shown in Figure 2, using a biased input structure (middle) both docking simulations were successful while using an unbiased input structure (bottom) only redocking succeeded. To understand how the covalent geometry influenced the performance of the simulations, the biased and unbiased input structures are analyzed. Figure 3 shows that for redocking the molecular conformations can be well superimposed and that no important deformation of the covalent geometry occurs upon binding. On the contrary, for cross-docking the geometry of the nitrogen in the central ring of the biased conformation is stretched and the unbiased structure significantly differs from the one minimized in 1hbv. These deformations can preclude the finding of the solution and were responsible for the cross-docking failure in experiments (3).

A careful analysis of the results shown in Figure 2 suggests that there is a correlation between the quality of the docking prediction and the convergence of multiple runs of the genetic algorithm. This is partially due to the fact that it is more difficult to dock highly flexible ligands for which the lack of convergence is a consequence of the large conformational space. In principle, the lack of convergence could then be used as a criterion for judging the quality of a docking prediction and as a valuable indicator in virtual high-throughput screening projects. To evaluate the reliability of a convergence-based criterion, only the cross-docking experiments of type (c) were considered, yielding a test sample of 20 docking simulations. The redocking simulations were discarded because they implicitly contain information of the bound complex. For each of the 20 docking simulations, convergence toward the lowest-energy conformation (not necessarily the experimental structure) in 10 FFLD runs with different random generator seeds was first determined (Fig. 4, top). The convergence values were then used to build the density plot shown in Figure 4 (bottom). The density plot is darker in the top right region, which indicates that convergence is a necessary condition for reliable predictions. Undoubtedly, the sample used for testing the convergence-based criterion is rather small (only 20 docking simulations) and therefore it is difficult to draw general conclusions. Nevertheless, the density plot shows a clear trend and implies that docking experi-

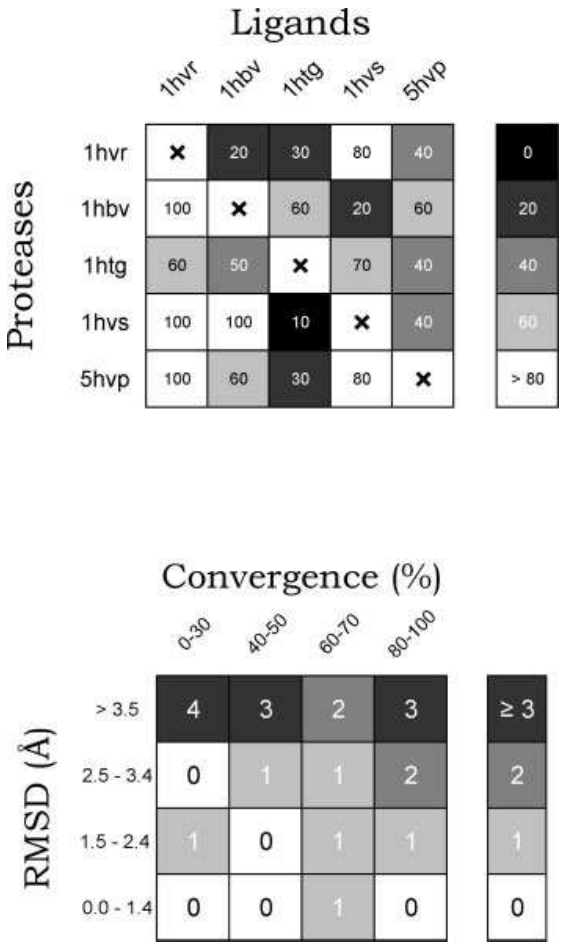


Figure 4. Evaluation of the convergence-based criterion proposed for judging the quality of the docking predictions. Top, convergence toward the lowest energy solution (not necessarily the experimental structure) in 10 docking runs is shown for the cross-docking experiments of type (3). Bottom, the density plot reports the frequency of finding a certain RMSD between the lowest-energy ligand pose and the experimental structure for a given amount of convergence; darker colors represent higher-frequency values.

ments with convergence lower than 60% may have not found the global minimum of the CHARMm22 energy surface and should be discarded during the analysis of a library screening project.

Convergence toward a unique binding mode in multiple runs of the genetic algorithm is a necessary but not sufficient condition for judging the quality of a prediction. It is not sufficient because an oversimplified energy function with a funnel-like profile together with a protein conformation that does not allow the reproduction of the crystal structure will yield convergence on a wrong binding mode.

Finally, to completely remove the geometric bias of the crystal structure a conformational search of the ligand 1hvs was performed by high-temperature molecular dynamics in the absence of the protein. The simulation was run for 2 ns at 400 K using the Berendsen thermostat and a distance-dependent dielectric function $[\varepsilon(r) = 4r]$. The final snapshot was minimized and its RMSD from the X-ray conformation after optimal overlap is 5.8 Å, which indicates that all of the information was lost and the initial conformation for docking was fully unbiased. Redocking was successful with a RMSD of 1.1 Å and a convergence of 80%. This result can not be generalized. On the contrary, it is likely that the majority of experiments (3) would deteriorate by using ligand covalent geometries without any memory of the crystal structure.

Results on Human α-Thrombin and the Estrogen Receptor β

To test the approach on binding sites with different physicochemical properties the same docking procedure was applied to human α-thrombin and the estrogen receptor β. Human α-thrombin presents an asymmetrical binding site with two hydrophobic pockets and a hydrophilic cavity specific for positively charged amino acids (Lys or Arg). Estrogen receptor β has a predominantly hydrophobic and almost completely buried binding site (see above). In both cases, starting from an unbiased and fully flexible conformation of the ligand [see type (3) docking experiments above] the SEED–FFLD approach was able to correctly reproduce the experimentally determined binding modes with RMSD smaller than 1 Å and convergences larger than 90%.

Judging Search Methods

To evaluate the performance of the hybrid search procedure it was compared with the GA of the original version of FFLD.[10]

For this purpose, unbiased redocking experiments were repeated without using the local optimizer during the evolution and with the same amount of energy evaluations. The simulations clearly show that a hybrid search is more efficient than GA because it always reaches a lower-energy conformation. The results of two redocking experiments carried out with both search methods are presented in Figure 5. For redocking 1hvr, a hybrid search is more efficient than the GA, especially at the beginning of the simulation, where the energy gap is large. At about 60% of the evolution the gap decreases and the performance of the two methods is comparable. For redocking 1hvs, the hybrid search performs better during the entire simulation and the energy gap increases until the end. Moreover, the standard deviation of the hybrid search evolutions (shown as error bars in Fig. 5) is larger, indicating that it is less prone to premature convergence than the GA.

The comparison shows that the local search accelerates convergence of the simulations and dramatically improves the quality of the docking predictions in case the conformational space of the ligand is large and its torsional degrees of freedom are strongly coupled (main-chain flexibility). This is mainly due to the fact that random perturbations of binary strings performed by the GA during the evolution correspond to radical jumps in the energy landscape and may be too large. On the contrary, the local optimizer is able to refine large perturbations due to crossover and mutations and leads to a better investigation of the energy landscape. The results of the present docking study suggest that hybrid search methods should be preferred to canonical genetic algorithms.

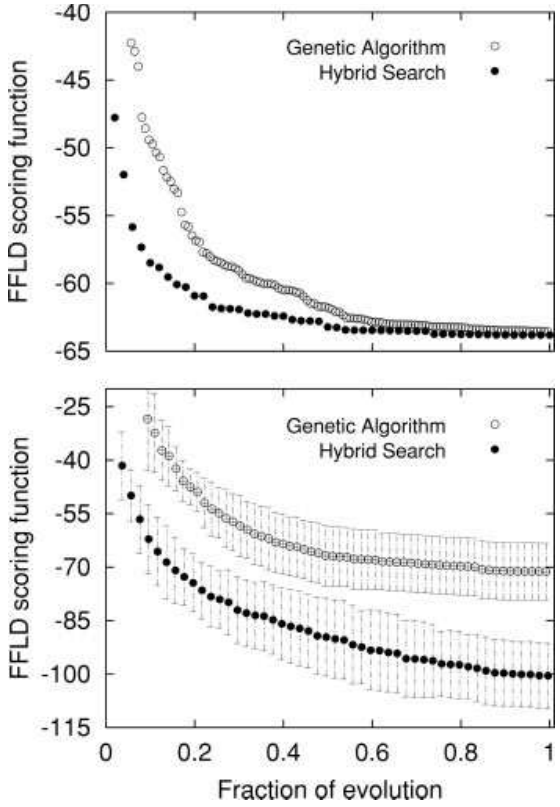


Figure 5. Evolution of the best individual of the population averaged over 10 docking runs for 2 experiments of type (3) using the same number of energy evaluations. Open and solid bullets indicate evolutions performed by genetic algorithm and hybrid search procedure, respectively. Redocking of the ligands 1hvr ($3 \cdot 10^5$ energy evaluations) and 1hvs ($1 \cdot 10^6$ energy evaluations) are shown from top to bottom. In the bottom plot, the vertical bars show the standard deviation computed over 10 docking runs.

Computational Requirement

All docking simulations were carried out on 1.6-GHz Athlon processors. For experiments of type (1), because of the limited amount of ligand flexibility fast docking calculations were performed, having a maximum number of 50 hybrid search cycles per run. The computational time required for a single docking varied from 5 to 19 min, yielding an average time of 12 min/run. For experiments of types (2) and (3), both carried out with full ligand flexibility, more extensive calculations were performed, using approximately 1 million energy evaluations per run. The computational time required for a single docking varied from 123–214 min, yielding an average time of 168 min/run. In these experiments, the number of energy evaluations per run was intentionally overestimated to make sure that the stochastic algorithm used for docking reached convergence in all cases. The computational requirements given above do not include docking of molecular fragments by SEED,[23] which was performed only once for each protein conformation.

Conclusions

Four main conclusions can be drawn from the docking results. First, a hybrid approach consisting of a local search and genetic algorithm significantly improves the quality of the SEED–FFLD docking predictions at a moderate additional computational cost. This is not a new finding[31] but simply provides further evidence with more stringent test cases, i.e., cross-docking experiments with highly flexible ligands.

Second, the quality of docking predictions depends on the degree of ligand flexibility and we suggest that validation of docking approaches should be always done with full dihedral flexibility of the ligands. In this respect, it would be interesting if the study of Österberg et al.[14] could be repeated without holding the main-chain of the peptidic inhibitors rigid.

Third, automatic approaches that sample only in dihedral space can give misdocked predictions if the covalent geometry of the ligand (i.e., its bond angles and lengths) is strained upon binding to its target. Therefore, a reliable validation of a docking approach should be performed without using any information on the conformation of the bound ligand, i.e., after a conformational search outside of the receptor. This was not done in previous works by us and others.[10,14,22,29,31,47] For docking a limited set of compounds, approaches that allow for full flexibility (including bond angles and lengths) of the ligands, albeit computationally expensive, should be preferred.[11,12]

Fourth, the docking results indicate that convergence toward the same docking solution in multiple runs of the genetic algorithm is a necessary (but not sufficient) condition for reliable predictions.

Acknowledgments

The authors are grateful to N. Budin, F. Dey and D. Huang for helpful discussions and to the referees for interesting comments and useful suggestions. The simulations were performed on a Beowulf cluster running Linux and the authors thank U. Haberthür and F. Rao for their help in setting up and maintaining the cluster. A. Widmer (Novartis Pharma, Basel) is thanked for providing the molecular modeling program Wit!P, which was used for visual analysis of the docking results. This work was supported by the Swiss National Competence Center in Structural Biology (NCCR).

References

1. Kuntz, I. D. Science 1992, 257, 1078.
2. Apostolakis, J.; Caflisch, A. Comb Chem High Throughput Screen 1999, 2, 91.
3. Glen, R. C.; Allen, S. C. Curr Med Chem 2003, 10, 763.
4. Walters, W. P.; Namchuk, M. Nature Rev Drug Discov 2003, 2, 259.
5. Gordon, E. M.; Barrett, R. W.; Dower, W. J.; Fodor, S. P. A.; Gallop, M. A. J Med Chem 1994, 37, 1385.
6. Zuckermann, R. N.; Martin, E. J.; Spellmeyer, D. C.; Stauber, G. B.; Shoemaker, K. R.; Kerr, J. M.; Figliozzi, G. M.; Goff, D. A.; Siani, M. A.; Simon, R. J.; Banville, S. C.; Brown, E. G.; Wang, L.; Richter, L. S.; Moos, W. H. J Med Chem 1994, 37, 2678.
7. Chang, Y.; Gray, N. S.; Rosania, G. R.; Sutherlin, D. P.; Kwon, S.; C.;

N. T.; Sarohia, R.; Leos, M.; Meijer, L.; Schultz, P. G. Chem Biol 1999, 6, 361.
8. Wang, J.; Kollman, P. A.; Kuntz, I. D. Proteins 1999, 36, 1.
9. Jones, G.; Willett, P.; Glen, R. C. J Mol Biol 1995, 245, 43.
10. Budin, N.; Majeux, N.; Caflisch, A. Biol Chem 2001, 382, 1365.
11. Given, J. A.; Gilson, M. K. Proteins 1998, 33, 475.
12. Apostolakis, J.; Plückthun, A.; Caflisch, A. J Comput Chem 1998, 19, 21.
13. Claussen, H.; Buning, C.; Rarey, M.; Lengauer, T. J Mol Biol 2001, 308, 377.
14. Österberg, F.; Morris, G. M.; Sanner, M. F.; Olson, A. J.; Goodsell, D. S. Proteins 2002, 46, 34.
15. Verkhivker, G. M.; Bouzida, D.; Gelhaar, D. K.; Rejto, P. A.; Freer, S. T.; Rose, P. W. Curr Opin Struct Biol 2002, 12, 197.
16. Wong, C. F.; McCammon, J. A. Annu Rev Pharmacol Toxicol 2003, 43, 31.
17. Kallblad, P.; Dean, P. M. J Mol Biol 2003, 326, 1651.
18. Halperin, I.; Ma, B. Y.; Wolfson, H.; Nussinov, R. Proteins 2002, 47, 409.
19. Taylor, R. D.; Jewsbury, P. J.; Essex, J. W. J Comp-Aided Mol Design 2002, 16, 151.
20. Lyne, P. D. Drug Discov Today 2002, 7, 1047.
21. Verkhivker, G. M.; Bouzida, D.; Gelhaar, D. K.; Rejto, P. A.; Arthurs, S.; Colson, A. B.; Freer, S. T.; Larson, V.; Luty, B. A.; Marrone, T.; Rose, P. W. J Comput-Aided Mol Design 2000, 14, 731.
22. Sotriffer, C. A.; Gohlke, H.; Klebe, G. J Med Chem 2002, 45, 1967.
23. Majeux, N.; Scarsi, M.; Apostolakis, J.; Ehrhardt, C.; Caflisch, A. Proteins 1999, 37, 88.
24. Scarsi, M.; Apostolakis, J.; Caflisch, A. J Phys Chem A 1997, 101, 8098.
25. Majeux, N.; Scarsi, M.; Caflisch, A. Proteins 2001, 42, 256.
26. Kabsch, W. Acta Crystallogr 1976, A32, 922.
27. Kearsley, S. K.; Smith, G. M. Tetrahedron Comp Meth 1990, 3, 615.
28. Klebe, G.; Mietzner, T.; Weber, F. J Comp-Aided Mol Design 1994, 8, 751.
29. Verkhivker, G. M.; Rejto, P. A.; Gelhaar, D. K.; Freer, S. T. Proteins 1996, 25, 342.
30. Gerber, P. R.; Müller, K. J Comp-Aided Mol Design 1995, 9, 251.
31. Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. J Comput Chem 1998, 19, 1639.
32. Budin, N.; Majeux, N.; Tenette-Souaille, C.; Caflisch, A. J Comput Chem 2001, 22, 1956.
33. Solis, F. J.; Wets, R. J. B. Math Oper Res 1981, 1, 19.
34. Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. J Comput Chem 1983, 4, 187.
35. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. Nucl Acids Res 2000, 28, 235.
36. Lam, P. Y. S.; Jadhav, P. K.; Eyermann, C. J.; Hodge, C. N.; Ru, Y.; Bacheler, L. T.; Meek, J. L.; Otto, M. J.; Rayner, M. M.; Wong, Y. N.; Chang, C. H.; Weber, P. C.; Jackson, D. A.; Sharpe, T. R.; Erickson-Viitanen, S. K. Science 1994, 263, 380.
37. Hoog, S. S.; Zhao, B.; Winborne, E.; Fischer, S.; Green, D. W.; DesJarlais, R. L.; Newlander, K. A.; Callahan, J. F.; Moore, M. L.; Huffman, W. F.; Abdel-Meguid, S. J Med Chem 1995, 38, 3246.
38. Jhoti, H.; Singh, O. M.; Weir, M. P.; Cooke, R.; Murray-Rust, P.; Wonacott, A. Biochemistry 1994, 33, 8417.
39. Baldwin, E. T.; Bhat, T. N.; Liu, B.; Pattabiraman, N.; Erickson, J. W. Nature Struct Biol 1995, 2, 244.
40. Fitzgerald, P. M. D.; Mckeever, B. M.; Vanmiddlesworth, J. F.; Springer, J. P.; Heimbach, J. C.; Leu, C. T.; Werber, W. K.; Dixon, R. A. F.; Darke, P. L. J Biol Chem 1990, 265, 14209.
41. Skrzypczak-Jankun, E.; Carperos, V. E.; Ravichandran, K. G.; Tulinsky, A.; Westbrook, M.; Maraganore, J. M. J Mol Biol 1991, 221, 1379.
42. Henke, B. R.; Consler, T. J.; Go, N.; Hale, R. L.; Hohman, D. R.; Jones, S. A.; Lu, A. T.; Moore, L. B.; Moore, J. T.; Orband-Miller, L. A.; Robinett, R. G.; Shearin, J.; Spearing, P. K.; Stewart, E. L.; Turnbull, P. S.; Weaver, S. L.; Williams, S. P.; Wisely, G. B.; Lambert, M. H. J Med Chem 2002, 45, 5492.
43. Piana, S.; Sebastiani, D.; Carloni, P.; Parrinello, M. J Am Chem Soc 2001, 123, 8730.
44. No, K.; Grant, J.; Scheraga, H. J Phys Chem 1990, 94, 4732.
45. No, K.; Grant, J.; Jhon, M.; Scheraga, H. J Phys Chem 1990, 94, 4740.
46. Tapparelli, C.; Metternich, R.; Ehrhardt, C.; Cook, N. S. Trends Pharmacol Sci 1993, 14, 366.
47. Gehlhaar, D. K.; Verkhivker, G. M.; Rejto, P. A.; Sherman, C. J.; Fogel, D. B.; Fogel, L. J.; Freer, S. T. Chem Biol 1995, 2, 317.
48. DeLano, W. The PyMOL Molecular Graphics System; DeLano Scientific: San Carlos, CA, 2002.

Chapter 4

Discovery of Cell-Permeable Non-Peptide Inhibitors of β-Secretase by High-Throughput Docking and Continuum Electrostatics Calculations

(Journal of Medicinal Chemistry 48, pp 5108-5111, 2005)

Discovery of Cell-Permeable Non-Peptide Inhibitors of β-Secretase by High-Throughput Docking and Continuum Electrostatics Calculations#

Danzhi Huang,[‡,†] Urs Lüthi,[§,†] Peter Kolb,[‡] Karin Edler,[‡,§] Marco Cecchini,[‡] Stephan Audetat,[‡,§] Alcide Barberis,[§] and Amedeo Caflisch*[,‡]

Department of Biochemistry, University of Zürich, Winterthurerstrasse 190, CH-8057 Zürich, Switzerland, and ESBATech AG, Wagistrasse 21, CH-8952 Zürich-Schlieren, Switzerland

Received May 27, 2005

Abstract: A fragment-based docking procedure followed by substructure search were used to identify active-site β-secretase inhibitors from a composite set of about 300 000 and a library of nearly 180 000 small molecules, respectively. EC_{50} values less than 10 μM were measured in at least one of two different mammalian cell-based assays for 12 of the 72 purchased compounds. In particular, the phenylureathiadiazole **2** and the diphenylurea derivative **3** are promising lead compounds for β-secretase inhibition.

Alzheimer's disease is the most common neurodegenerative disease and accounts for the majority of the dementia diagnosed after the age of 60.[1] Amyloid plaques, which are found in the post-mortem brain of Alzheimer's disease patients,[2] consist mainly of fibrillar aggregates of the Aβ peptide, a proteolytic cleavage product of the β-amyloid precursor protein (APP). Two enzymes, γ- and β-secretase (β-site APP cleaving enzyme, or BACE-1), are responsible for the sequential processing of APP.[3] Although it is not clear whether the plaques or oligomeric prefibrillar species are responsible for neuronal loss and dementia,[4] the pepsin-like aspartic protease BACE-1 has become one of the major Alzheimer's disease targets.[1,5] BACE-1 is a very difficult target as is witnessed by the very small number of known nonpeptidic inhibitors.[1,5−7] Moreover, not a single BACE-1 inhibitor was found in a library containing more than 1800 renin inhibitors,[8] despite the fact that both BACE-1 and renin are pepsin-like enzymes. In addition, a single molecule (1,3,5-trisubstituted benzene) emerged as a BACE-1 inhibitor from a multimillion compound library submitted to a high-throughput screening campaign.[9]

Here, we report the identification of a dozen BACE-1 inhibitors with a common phenylurea scaffold by our in silico screening approach that consists of four steps (details of the methods are in Supporting Information). First, each molecule is automatically decomposed into rigid fragments by the program DAIM (decomposition and identification of molecules; P. Kolb and A. Caflisch, manuscript in preparation). In a second step the fragments are docked into the rigid binding site by the program SEED,[10,11] which approximates solvation effects by continuum electrostatics.[12] As an improvement with respect to previous versions of SEED,[10,11,13] the screened electrostatic interaction and fragment desolvation energy were evaluated using an empirical correction of the Coulomb field approximation, i.e., eq 8 of ref 14. In the third step the optimal SEED binding modes of the fragments are then used as binding site descriptors to guide the placement of the flexible molecules by the docking program FFLD (fragment-based flexible ligand docking), which is based on a genetic algorithm.[13,15] The most favorable FFLD binding modes are further minimized in the rigid protein using the CHARMM program.[16]

The final step of our approach is the evaluation of the binding free energy with solvation effects,[17] which is an essential element of the in silico screening procedure. Computer-aided approaches for docking libraries of small molecules into proteins of known structure require fast and accurate methods for the evaluation of binding free energies.[18−22] Rigorous approaches to evaluate relative binding affinities such as free energy perturbation and thermodynamic integration have sampling and convergence problems that prevent them from being used routinely.[23] Moreover, it is very difficult to handle large 2D structural diversity between ligands, e.g., in the case of completely different core structures.[18] Several semiempirical methods based on linear approximations to the free energy have been introduced and used with success.[22] A decade ago Åqvist and co-workers proposed the LIE (linear interaction energy) method to calculate free energies of binding by averaging interaction energies from molecular dynamics (MD) simulations of the ligand and the ligand/protein complex.[24,25] To improve efficiency, which is essential for evaluating large libraries of compounds, we have replaced the MD sampling with a simple energy minimization and combined the LIE method with a rigorous treatment of continuum electrostatics, i.e., numerical solution of the Poisson equation by the finite-difference technique.[26] The modified LIE approach, termed LIECE where the last two letters stand for continuum electrostatics, was shown to have an accuracy in the binding energy prediction of about 1 kcal/mol for a set of 13 and 29 peptidic inhibitors of BACE-1 and HIV-1 aspartic protease, respectively.[17] It was also shown that a LIECE model parametrized on HIV-1 aspartic protease is not transferable to BACE-1 and vice versa.[17] Hence, in general the LIECE approach cannot be used in virtual screening against a target for which no inhibitor is known. On the other hand, a recent application to three different kinases indicates transferability of the LIECE parameters (Huang, Kolb, and Caflisch, unpublished results).

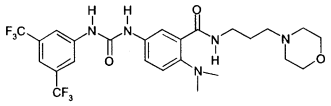
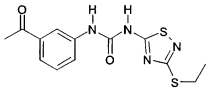
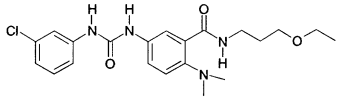
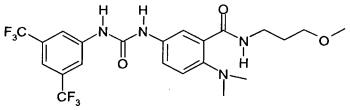
Initially, about 300 000 molecules with at least one hydroxyl group were selected from a collection of chemical libraries containing about six million compounds. The in silico screening of these 300 000 molecules, i.e., docking and LIECE energy evaluation, took about 10 days on a Beowulf cluster of 100 1.8-GHz Opteron

This paper is dedicated to Martin Karplus on the occasion of his 75th birthday.

* To whom correspondence should be addressed. Phone: (+41 44) 635 55 21. Fax: (+41 44) 635 68 62. E-mail: caflisch@bioc.unizh.ch.
‡ University of Zürich.
§ ESBATech AG.
† D.H. and U.L. contributed equally to this work.

10.1021/jm050499d CCC: $30.25
Published on Web 07/08/2005

Table 1

Compounds	Structures	MW (g mol^{-1})	BACE-1[a] IC_{50} (μM)	Abeta(sw)[b] EC_{50} (μM)	SEAP[c] EC_{50} (μM)	Cytotoxic[d] CC_{50} (μM)	LIECE[e] K_i (μM)
1		561.5	57.8 ± 10.3	3.0 ± 0.6	3.5 ± 0.7	12.5	8.1
2		322.4	>25 [f]	2.6 ± 0.9	11.4 ± 0.9	22.3	28.8
3		418.9	97.0 ± 21.4	2.6 ± 1.1	23.3 ± 8.1 [g]	11.1	9.8
4		506.5	283.9 ± 36.8	3.2 ± 0.2	27.0 ± 9.5 [g]	24.6	9.4

[a] The BACE-1 fluorescence resonance energy transfer assay kit was purchased from PanVera (Madison, WI; no. P2985). BACE-1 activity assays were carried out according to the manufacturer's instructions. Average value and standard deviation are from three independent experiments. [b] Cell-based assay.[28] Average value and standard deviation are from three independent experiments. [c] Cell-based assay.[29] Average value and standard deviation are from three independent experiments. [d] Cytotoxic concentration.[33] [e] See ref 17. [f] Interference at concentrations higher than 25 μM. [g] Percentage inhibition at 3 μM.

CPUs. The rigid conformation of BACE-1 from its complex with the OM00-3 inhibitor (PDB code 1m4h [27]) was used for the docking. Interestingly, only 10 compounds had a LIECE-predicted affinity in the high-nanomolar range and most of them were phenylurea derivatives with the two NH groups involved in hydrogen bonds with one of the two catalytic aspartates. Unfortunately, these 10 compounds were no longer available from the original vendor. Therefore, we decided to select from the six million molecule collection all of the nearly 32 000 compounds with a phenylurea moiety, i.e., those with (only 1233 molecules) and without a hydroxyl group. These nearly 32 000 compounds were docked; the poses with the most favorable FFLD energy were further minimized by CHARMM,[16] and the energetically most favorable 50 000 poses (8558 different molecules) were evaluated by the LIECE approach.[17] The LIECE binding energy evaluation was performed in two steps using first a grid spacing of 1.0 Å in the finite-difference Poisson calculation followed by a more accurate calculation with a grid spacing of 0.3 Å for the best 2000 poses. The two-step LIECE procedure required about 20 h on the Beowulf cluster of 100 CPUs. Upon visual inspection of the top 200 poses (131 different molecules), 10 compounds were purchased and tested in an enzymatic assay with purified BACE-1 and in two cell-based assays. We first tested the cellular activity of the selected compounds by measuring Aβ peptide secretion.[28] To confirm BACE-1 inhibition in an additional mammalian cellular assay we established the so-called SEAP (secreted alkaline phosphatase) system. For this system, HEK 293 cells were transfected with a SEAP-APP fusion protein bearing the SEAP enzyme moiety localized in the topologically extracellular space, such as ER/Golgi lumen and endosomes, or also at the cell surface.[29] This protein is anchored to cellular membranes via a portion of APP harboring the Swedish mutation at the β-site and the K612V mutation at the α-site. Endogenous β-secretase activity causes liberation and subsequent secretion of the SEAP enzyme, whose activity in the supernatant is measured via a chemiluminescent read-out. In this way, the diphenylurea derivative **1** (Table 1) was identified as a low-micromolar inhibitor of BACE-1. Two of the remaining nine compounds showed low-micromolar activity in at least one of the two cell-based assays and the enzymatic assay (data not shown).

An essentially identical screening approach based on FFLD docking and LIECE postprocessing was applied to the 2476 compounds in a protease-focused chemical library. Intriguingly, seven among the 20 compounds with the most favorable LIECE-predicted affinity had a phenylurea scaffold. These 20 compounds were purchased and tested. The phenylurea derivative **2** showed low-micromolar activity in two different mammalian cell-based assays (Table 1). One of the remaining 19 compounds showed low-micromolar activity in both cell-based assays and an IC_{50} of 490 μM in the enzymatic assay (data not shown).

In a third in silico screening, 391 compounds from a libary of about 180 000 small molecules were first selected by similarity search using the phenylurea scaffold. After the FFLD docking and LIECE postprocessing, 42 compounds were purchased and tested. At 10 μM, 38 of the 42 compounds showed more than 20% inhibition in at least one of the two cell-based assays. Moreover, 10 of them have $EC_{50} < 10$ μM in the Abeta-(sw) assay. The two most potent BACE-1 inhibitors obtained by the similarity search and docking approach (**3** and **4**) are shown in Table 1. Despite its smaller size, **3** is as active as **4** in the two cell-based assays and a factor of about 3 more active in the enzymatic test.

It is interesting to compare **2** with the known non-peptidic inhibitors of BACE-1 which, as mentioned above, are rare.[1,5] A series of hydroxyethylamine deriva-

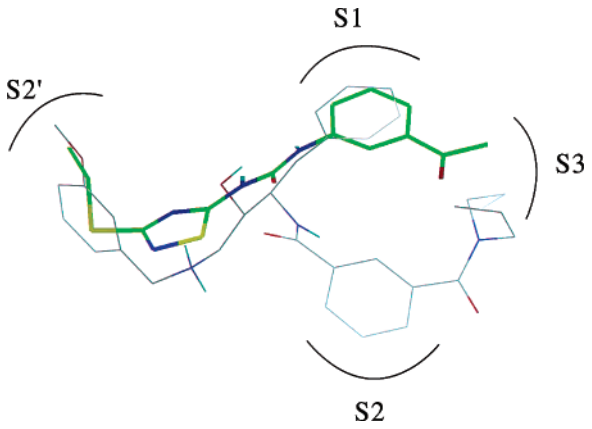


Figure 1. Superposition of a known nanomolar inhibitor of BACE-1[31] (thin lines and carbon atoms in gray) and **2** (thick lines with carbon atoms in green) in one of its two possible orientations obtained by docking in the flexible binding site. The C_α atoms of BACE-1 were used for the structural alignment.

tives of an isophthalamide scaffold have been shown to have nanomolar affinity by enzymatic[30−32] and cell-based assays.[32] The crystal structures of two of these inhibitors in complex with BACE-1 show that they have a very similar binding mode despite the different stereochemistry at the hydroxyl group.[31,32] In the catalytic site, the hydroxyl functionality and the protonated secondary amino group are involved in hydrogen bonds with the side chain of the catalytic Asp32 and Asp228, respectively. Moreover, the benzyl functionality close to the hydroxyl group of the two inhibitors occupies the S1 pocket in both complexes. The molecular weight of the hydroxyethylamine compounds (MW = 531 g mol^{-1} (ref 30) and MW = 579 g mol^{-1} for **3** (ref 32)) is larger than the one of compound **2** reported here (MW = 322 g mol^{-1}). Furthermore, the binding mode is different except for the phenyl group of the inhibitor **2** which occupies the S1 pocket (Figure 1) and overlaps with the corresponding ring of the benzyl functionality of the hydroxyethylamine inhibitors. Because of the small size and rather symmetric overall shape of **2**, we decided to perform minimization in the flexible binding site (library docking had been performed in the rigid protein) starting from the two end-to-end flipped orientations obtained by the FFLD docking. An alternative binding mode of inhibitor **2** is observed upon minimization in the flexible binding site with protonated Asp32 (instead of Asp228, which was protonated in all other calculations). In the alternative binding mode, the two NH groups of the urea scaffold are involved in hydrogen bonds with Asp228 (instead of Asp32), but the overall orientation is flipped end-to-end such that the ethylthioether functionality and the phenyl group occupy the S1 and S1′ pockets, respectively (Figure 2). It is not possible to apply LIECE to evaluate the two different binding modes because the LIECE approach requires a single protein conformation as reference state. Hence, the CHARMM in vacuo interaction energy supplemented by the finite-difference Poisson solvation was calculated for both binding modes, but the preferred orientation cannot be determined because the energy difference of 2.4 kcal/mol is within the limited accuracy of the estimation due to the flexible protein treatment. It is important to note that **2** has only two rotatable

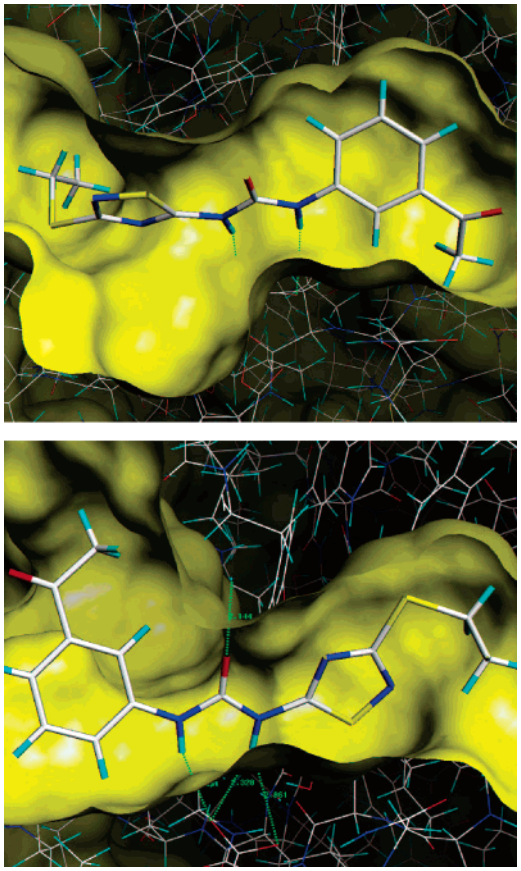


Figure 2. Two possible binding modes of **2** in the BACE-1 active site. Hydrogen bonds are shown by green dotted lines. The binding mode in the top picture corresponds to the one of Figure 1.

bonds. Its limited flexibility and the marginal loss of entropy upon binding are consistent with its rather high binding affinity given the small size.

In conclusion, high-throughput docking into the BACE-1 active site and continuum electrostatics calculations were used to select for experimental testing 72 compounds from an initial set of about 500 000. Fifty-nine of these 72 compounds are phenylurea derivatives. Twelve of the 72 compounds inhibit BACE-1 in at least one of two different mammalian cell-based assays at concentration values less than 10 μM. It is important to note that for almost all of the 12 compounds, for which an EC_{50} value could be measured, the discrepancies between LIECE-predicted affinity and the experimental value is within the LIECE accuracy of about 1 kcal/mol.[17] Given their very small size, the phenylureathiadiazole **2** (MW = 322 g mol^{-1}) and diphenylurea derivative **3** (MW = 419 g mol^{-1}) may serve as starting points for further optimization to evaluate their therapeutic potential for Alzheimer's disease.

Acknowledgment. We are grateful to Dr. Nicolas Majeux and Fabian Dey for interesting comments and useful suggestions. The calculations were performed on Matterhorn, a Beowulf Linux cluster at the Informatikdienste of the University of Zurich, and we thank C. Bolliger, Dr. T. Steenbock, and Dr. A. Godknecht for installing and maintaining the Linux cluster. This work was supported by a KTI grant to A.C.

Supporting Information Available: Details on computation approach and experimental tests. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Citron, M. β-Secretase inhibition for the treatment of Alzheimer's disease: promise and challenge. *Trends Pharmacol. Sci.* **2004**, *25*, 92−97.
(2) Selkoe, D. J. Translating cell biology into therapeutic advances in Alzheimer's disease. *Nature* **1999**, *399*, A23−A31.
(3) Lin, X.; Koelsch, G.; Wu, S.; Downs, D.; Dashti, A.; et al. Human aspartic protease memapsin 2 cleaves the β-secretase site of β-amyloid precursor protein. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 1456−1460.
(4) Petkova, A. T.; Leapman, R. D.; Guo, Z.; Yau, W. M.; Mattson, M. P.; et al. Self-propagating, molecular-level polymorphism in Alzheimer's β-amyloid fibrils. *Science* **2005**, *307*, 262−265.
(5) Cumming, J. N.; Iserloh, U.; Kennedy, M. E. Design and development of BACE-1 inhibitors. *Curr. Opin. Drug Discovery Dev.* **2004**, *7*, 536−556.
(6) Roggo, S. Inhibition of BACE, a promising approach to Alzheimer's disease therapy. *Curr. Top. Med. Chem.* **2002**, *2*, 359−370.
(7) Middendorp, O.; Lüthi, U.; Hausch, F.; Barberis, A. Searching for the most effective screening system to identify cell-active inhibitors of β-secretase. *Biol. Chem.* **2004**, *385*, 481−485.
(8) Grüninger-Leitch, F.; Schlatter, D.; Küng, E.; Nelböck, P.; Döbeli, H. Substrate and inhibitor profile of BACE and comparison with other mammalian aspartic proteases. *J. Biol. Chem.* **2002**, *277*, 4687−4693.
(9) Coburn, C. A.; Stachel, S. J.; Li, Y. M.; Rush, D. M.; Steele, T. G.; et al. Identification of a small molecule nonpeptide active site β-secretase inhibitor that displays a nontraditional binding mode for aspartyl proteases. *J. Med. Chem.* **2004**, *47*, 6117−6119.
(10) Majeux, N.; Scarsi, M.; Apostolakis, J.; Ehrhardt, C.; Caflisch, A. Exhaustive docking of molecular fragments on protein binding sites with electrostatic solvation. *Proteins: Struct., Funct., Genet.* **1999**, *37*, 88−105.
(11) Majeux, N.; Scarsi, M.; Caflisch, A. Efficient electrostatic solvation model for protein-fragment docking. *Proteins: Struct., Funct., Genet.* **2001**, *42*, 256−268.
(12) Scarsi, M.; Apostolakis, J.; Caflisch, A. Continuum electrostatic energies of macromolecules in aqueous solutions. *J. Phys. Chem. A* **1997**, *101*, 8098−8106.
(13) Cecchini, M.; Kolb, P.; Majeux, N.; Caflisch, A. Automated docking of higly flexible ligands by genetic algorithms: A critical assessment. *J. Comput. Chem.* **2004**, *25*, 412−422.
(14) Lee, M. S.; Salsbury, F. R.; Brooks, C. L., III. Novel generalized Born methods. *J. Chem. Phys.* **2002**, *116*, 10606−10614.
(15) Budin, N.; Majeux, N.; Caflisch, A. Fragment-based flexible ligand docking by evolutionary opimization. *Biol. Chem.* **2001**, *382*, 1365−1372.
(16) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; et al. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **1983**, *4*, 187−217.
(17) Huang, D.; Caflisch, A. Efficient evaluation of binding free energy using continuum electrostatic solvation. *J. Med. Chem.* **2004**, *47*, 5791−5797.
(18) Jorgensen, W. L. The many roles of computation in drug discovery. *Science* **2004**, *303*, 1813−1818.
(19) Glen, R. C.; Allen, S. C. Ligand−protein docking: Cancer research at the interface between biology and chemistry. *Curr. Med. Chem.* **2003**, *10*, 763−777.
(20) Walters, W. P.; Namchuk, M. Designing screens: How to make your hits a hit. *Nat. Rev. Drug Discovery* **2003**, *2*, 259−266.
(21) Doman, T. N.; McGovern, S. L.; Witherbee, B. J.; Kasten, T. P.; Kurumbail, R.; et al. Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1B. *J. Med. Chem.* **2002**, *45*, 2213−2221.
(22) Apostolakis, J.; Caflisch, A. Computational ligand design. *Comb. Chem. High Throughput Screening* **1999**, *2*, 91−104.
(23) Kollman, P. A. Free energy calculations: Applications to chemical and biochemical phenomena. *Chem. Rev.* **1993**, *93*, 2395−2417.
(24) Åqvist, J.; Medina, C.; Samuelsson, J.-E. A new method for predicting binding affinity in computer-aided drug design. *Protein Eng.* **1994**, *7*, 385−391.
(25) Hansson, T.; Åqvist, J. Estimation of binding free energies for HIV proteinase inhibitors by molecular dynamics simulations. *Protein Eng.* **1995**, *8*, 1137−1144.
(26) Warwicker, J.; Watson, H. C. Calculation of the electric potential in the active site cleft due to α-helix dipoles. *J. Mol. Biol.* **1982**, *157*, 671−679.
(27) Hong, L.; Turner, R. T.; Koelsch, G.; Shin, D.; Ghosh, A. K.; et al. Crystal structure of memapsin 2 (β-secretase) in complex with an inhibitor OM00-3. *Biochemistry* **2002**, *41*, 10963−10967.
(28) Dovey, H. R.; Suomensaari-Chrysler, S.; Lieberburg, L.; Sinha, S.; Keim, P. S. Cells with a familial Alzheimer's disease mutation produce authentic beta-peptide. *NeuroReport* **1993**, *4*, 1039−1042.
(29) Oh, M.; Kim, S. Y.; Oh, Y. S.; Choi, D.; Sin, H. J.; et al. Cell-based assay for beta-secretase activity. *Anal. Biochem.* **2003**, *323*, 7−11.
(30) Maillard, M.; Hom, C.; Gailunas, A.; Jagodzinska, B.; Fang, L. Y.; et al. Preparation of substituted amines to treat alzheimer's disease. WO-0202512, 2002.
(31) Patel, S.; Vuillard, L.; Cleasby, A.; Murray, C. W.; Yon, J. Apo and inhibitor complex structures of BACE (β-secretase). *J. Mol. Biol.* **2004**, *343*, 407−416.
(32) Stachel, S. J.; Coburn, C. A.; Steele, T. G.; Jones, K. G.; Loutzenhiser, E. F.; et al. Structure-based design of potent and selective cell-permeable inhibitors of human β-secretase (BACE-1). *J. Med. Chem.* **2004**, *47*, 6447−6450.
(33) Scudiero, D. A.; Shoemaker, R. H.; Paull, K. D.; Monks, A.; Tierney, S.; et al. Evaluation of a soluble tetrazolium/formazan assay for cell growth and drug sensitivity in culture using human and other tumor cell lines. *Cancer Res.* **1988**, *48*, 4827−4833.

JM050499D

Chapter 5

In Silico Discovery of β-Secretase Inhibitors

(Journal of the American Chemical Society 128, pp 5436-5443, 2006)

J|A|C|S
ARTICLES
Published on Web 03/30/2006

In Silico Discovery of β-Secretase Inhibitors

Danzhi Huang,[†] Urs Lüthi,[‡] Peter Kolb,[†] Marco Cecchini,[†] Alcide Barberis,[‡] and Amedeo Caflisch*,[†]

Contribution from the Department of Biochemistry, University of Zürich, Winterthurerstrasse 190, CH-8057 Zürich, Switzerland, and ESBATech AG, Wagistrasse 21, CH-8952 Zürich-Schlieren, Switzerland

Received October 26, 2005; E-mail: caflisch@bioc.unizh.ch

Abstract: Alzheimer's disease, the most common amyloid-associated disorder, accounts for the majority of the dementia diagnosed after the age of 60. The cleavage of the β-amyloid precursor protein is initiated by β-secretase (BACE-1), a membrane-bound aspartic protease, which has emerged as an important but difficult protein target. Here, an in silico screening approach consisting of fragment-based docking, ligand conformational search by a genetic algorithm, and evaluation of free energy of binding was used to identify low-molecular-weight inhibitors of BACE-1. More than 300 000 small molecules were docked and about 15 000 prioritized according to a linear interaction energy model with evaluation of solvation by continuum electrostatics. Eighty-eight compounds were tested in vitro, and 10 of them showed an IC_{50} value lower than 100 μM in a BACE-1 enzymatic assay. Interestingly, the 10 active compounds shared a triazine scaffold. Moreover, four of them were active in an assay with mammalian cells ($EC_{50} < 20$ μM), indicating that they are cell-permeable. Therefore, these triazine derivatives are very promising lead candidates for BACE-1 inhibition. The discoveries of this series and two other series of nonpeptidic BACE-1 inhibitors demonstrate the usefulness of our in silico high-throughput screening approach.

Introduction

Insoluble, extracellular amyloid plaques, a histopathological hallmark in the post-mortem brain of Alzheimer's disease patients,[1] consist mainly of fibrillar aggregates of the amyloid-β (Aβ) peptide, which is a proteolytic cleavage product of the β-amyloid precursor protein (APP). Two enzymes, γ- and β-secretase (β-site APP cleaving enzyme, or BACE-1), are responsible for the sequential processing of APP.[2] Genetic deletion of BACE-1 in mice has been shown to abolish β-amyloid formation with an otherwise normal, i.e., healthy, phenotype.[3] Although there is no definitive evidence whether the plaques or oligomeric prefibrillar species are responsible for neuronal loss and dementia,[4] the pepsin-like aspartic protease BACE-1 is considered an important target for the development of small-molecule inhibitors to fight Alzheimer's disease.[5,6] The relatively small number of known nonpeptide inhibitors indicates that BACE-1 is not an easy target to block.[5−8] In fact, not a single BACE-1 inhibitor was found in a library containing more than 1800 renin inhibitors,[9] despite the fact that both BACE-1 and renin are pepsin-like enzymes. Furthermore, only a single molecule (1,3,5-trisubstituted benzene) emerged as BACE-1 inhibitor from a multimillion compound library submitted to a high-throughput in vitro screening campaign.[10] It is also important to note that the recently reported peptidomimetics with low nanomolar affinity in BACE-1 enzymatic assays are not active in cell-based assays because of limited penetration across cell membranes.[11] Here, we report the successful application of our in silico high-throughput docking approach in the screening of more than 300 000 existing compounds, which has resulted in the discovery of a series of nonpeptide BACE-1 inhibitors with a common (1,3,5-triazin-2-yl)hydrazone scaffold. The fragment-based docking procedure, which takes into account electrostatic solvation, shows a high hit rate and generates few false positives. Most notably, the combination of in silico screening with validation by enzymatic and cell-based assays has led to the identification of several molecules with excellent potential as lead compounds against BACE-1.

Methods

The essential elements of our in silico screening are a fragment-based docking procedure and an efficient evaluation of binding free energy with electrostatic solvation. The latter is presented first because of its importance for ranking compounds, which is the most challenging part of the in silico approach.

[†] University of Zürich.
[‡] ESBATech AG.

(1) Selkoe, D. J. *Nature* **1999**, *399*, A23−A31.
(2) Lin, X.; Koelsch, G.; Wu, S.; Downs, D.; Dashti, A.; Tang, J. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 1456−1460.
(3) Luo, Y.; et al. *Nat. Neurosci.* **2001**, *4*, 231−232.
(4) Petkova, A. T.; Leapman, R. D.; Guo, Z.; Yau, W. M.; Mattson, M. P.; Tycko, R. *Science* **2005**, *307*, 262−265.
(5) Citron, M. *Trends Pharmacol. Sci.* **2004**, *25*, 92−97.
(6) Cumming, J. N.; Iserloh, U.; Kennedy, M. E. *Curr. Opin. Drug. Discovery Dev.* **2004**, *7*, 536−556.
(7) Roggo, S. *Curr. Top. Med. Chem.* **2002**, *2*, 359−370.
(8) Middendorp, O.; Lüthi, U.; Hausch, F.; Barberis, A. *Biol. Chem.* **2004**, *385*, 481−485.
(9) Grüninger-Leitch, F.; Schlatter, D.; Küng, E.; Nelböck, P.; Döbeli, H. *J. Biol. Chem.* **2002**, *277*, 4687−4693.
(10) Coburn, C. A.; Stachel, S. J.; Li, Y. M.; Rush, D. M.; Steele, T. G.; Chen-Dodson, E.; Holloway, M. K.; Munshi, S.; Simon, A. J.; Kuo, L.; Vacca, J. P. *J. Med. Chem.* **2004**, *47*, 6117−6119.
(11) Hanessian, S.; et al. *J. Med. Chem.* **2005**, *48*, 5175−5190.

10.1021/ja0573108 CCC: $33.50

Evaluation of Binding Free Energy with LIECE. The linear interaction energy with continuum electrostatics (LIECE) approach was recently reported elsewhere.[12] Here, only a brief overview of the method is presented, while the in-depth validation for BACE-1 is given in the Results and Discussion. The essential aspect of the linear interaction energy (LIE) method is that the free energy of binding can be calculated by considering only the end points of the thermodynamic cycle of ligand binding, i.e., bound and free states. For this purpose, Åqvist and co-workers proposed to calculate average values of interaction energies from molecular dynamics (MD) simulations of the isolated ligand and the ligand/protein complex.[13,14] They approximated the free energy of binding by

$$\Delta G_{\text{bind}} = {}^1/_2 \left(\langle E^{\text{elec}} \rangle_{\text{bound}} - \langle E^{\text{elec}} \rangle_{\text{free}} \right) + \alpha \left(\langle E^{\text{vdW}} \rangle_{\text{bound}} - \langle E^{\text{vdW}} \rangle_{\text{free}} \right) \quad (1)$$

where E^{elec} and E^{vdW} are the electrostatic and van der Waals interaction energies between the ligand and its surroundings. The surroundings are either the solvent (free) or the solvated ligand/protein complex (bound). The $\langle\ \rangle$ denotes an ensemble average sampled over a molecular dynamics (MD)[13] or Monte Carlo[15] trajectory, and the parameter α is determined empirically.[13] The LIE method is faster than rigorous free energy perturbation techniques and has been successfully applied in the design of a series of inhibitors of the malarial aspartic proteases Plm I and II.[16] Yet, LIE cannot be used for high-throughput docking because of its computational requirements (about 1 day for each compound). Therefore, we have replaced the MD sampling with a simple energy minimization and combined the LIE method with a rigorous treatment of solvation within the continuum electrostatics (CE) approximation,[12] i.e., the numerical solution of the Poisson equation by the finite-difference technique.[17] The LIECE approach is about 2 orders of magnitude faster than previous LIE methods and shows a similar precision on the targets tested. In fact, a predictive accuracy of about 1.0 kcal/mol was observed for 13 and 29 peptidic inhibitors of BACE-1 and HIV-1 protease, respectively.[12]

Preparation of the BACE-1 Structure. The X-ray structure of BACE-1 from its complex with a nanomolar peptidic inhibitor (PDB code 1M4H[18]) was used for the in silico screening because BACE-1/nonpeptide inhibitor structures were not available when this work was initiated. The side chain of Asp228 was protonated[12] (see also below), while all other Asp and Glu side chains were considered negatively charged and the Lys and Arg positively charged. Further details on the protein preparation for docking can be found in the Supporting Information.

Preparation of the Compound Libraries. Two unrelated libraries were screened in silico. The first contains about 10 000 molecules with an average molecular weight of 497.3 ± 42.8 g/mol (Chemdiv Inc., 2002). The second library is a subset of about 300 000 molecules (424.9 ± 71.4 g/mol) selected from a collection of chemical libraries of about six million compounds (Chemnavigator Inc., 2004). For this selection, the size and physicochemical character of the substrate binding site were taken into account by filtering out compounds with molecular weight smaller than 200 g/mol or larger than 700 g/mol, and molecules without at least one hydrogen bond donor and acceptor. The 2D-to-3D conversion was performed using CORINA.[19] This step was followed by the determination of the protonation state and hydrogen coordinates generation with BABEL,[20] the assignment of CHARMm atom types[21] and partial charges,[22,23] and energy minimization with a distance-dependent dielectric function.

High-Throughput Fragment-Based Docking. The library-docking approach consists of four consecutive steps: (1) decomposition of each molecule of the library into mainly rigid fragments, (2) fragment docking with evaluation of electrostatic solvation, (3) flexible docking of each molecule of the library using the position and orientation of its fragments as anchors, and (4) LIECE evaluation of the binding free energy for the best poses. The first three steps are performed by in-house-developed computer programs, while CHARMM[24] is used for the energy minimization and finite-difference Poisson calculations in the fourth step. The main aspects of the docking approach are illustrated in the four following subsections, while the details are given in the Supporting Information.

(1) Decomposition of Library Compounds into Fragments. The decomposition of a molecule into mainly rigid substructures and the selection of the three anchor fragments for docking are performed by the program DAIM (Decomposition And Identification of Molecules, P. Kolb and A. Caflisch, unpublished results). The major rules are listed in the Supporting Information. The decomposition generates mainly rigid fragments which can be docked very efficiently (see below).

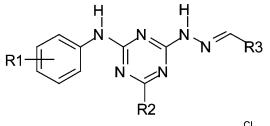
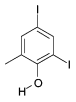
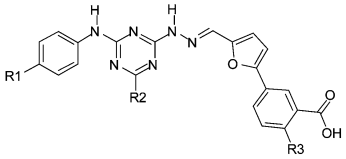
(2) Fragment Docking with Evaluation of Electrostatic Solvation. The docking approach implemented in the program SEED determines optimal positions and orientations of small to medium-size molecular fragments in the binding site of a protein.[25,26] Apolar fragments are docked into hydrophobic regions of the receptor, while polar fragments are positioned such that at least one intermolecular hydrogen bond is formed. Each fragment is placed at several thousand different positions with multiple orientations (for a total of in the order of 10^6 conformations), and the binding energy is estimated whenever severe clashes are not present (usually about 10^5 conformations). The binding energy is the sum of the van der Waals interaction and the electrostatic energy. The latter consists of screened receptor−fragment interaction, as well as values of receptor and fragment desolvation.[27]

(3) Flexible Docking of Library Compounds. The flexible-ligand docking approach FFLD uses a genetic algorithm and a very efficient but approximate scoring function.[28,29] FFLD requires three not necessarily different fragments to place a flexible ligand unambiguously in the binding site, e.g., the fluorobenzene, piperidine, and phenol of compound **5** (Table 1). Solvation effects are implicitly accounted for as the binding modes of the fragments are determined with electrostatic solvation in SEED. Each molecule was docked by three independent FFLD runs using a population of 100 members for each run and different initial values for the random number generator.

(4) Clustering and LIECE Binding Energy Evaluation. For each compound, the best 150 FFLD poses (50 poses from each FFLD run) were clustered by using a leader algorithm with a similarity cutoff of 0.7.[25,30] The representative of each cluster was selected for further

(12) Huang, D.; Caflisch, A. *J. Med. Chem.* **2004**, *47*, 5791−5797.
(13) Åqvist, J.; Medina, C.; Samuelsson, J.-E. *Protein Eng.* **1994**, *7*, 385−391.
(14) Hansson, T.; Åqvist, J. *Protein Eng.* **1995**, *8*, 1137−1144.
(15) Jones-Hertzog, D. K.; Jorgensen, W. H. *J. Med. Chem.* **1996**, *40*, 1539−1549.
(16) Ersmark, K.; Nervall, M.; Hamelink, E.; Janka, L. K.; Clemente, J. C.; Dunn, B. M.; Blackman, M. J.; Samuelsson, B.; Åqvist, J.; Hallberg, A. *J. Med. Chem.* **2005**, *48*, 6090−6106.
(17) Warwicker, J.; Watson, H. C. *J. Mol. Biol.* **1982**, *157*, 671−679.
(18) Hong, L.; Turner, R. T.; Koelsch, G.; Shin, D.; Ghosh, A. K.; Tang, J. *Biochemistry* **2002**, *41*, 10963−10967.
(19) Sadowski, J.; Gasteiger, J. *Chem. Rev.* **1993**, *93*, 2567−2581.
(20) Stahl, M. T. *Drug Discovery Today* **2005**, *10*, 219−222.
(21) Momany, F.; Rone, R. *J. Comput. Chem.* **1992**, *13*, 888−900.
(22) No, K.; Grant, J.; Scheraga, H. *J. Phys. Chem.* **1990**, *94*, 4732−4739.
(23) No, K.; Grant, J.; Jhon, M.; Scheraga, H. *J. Phys. Chem.* **1990**, *94*, 4740−4746.
(24) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187−217.
(25) Majeux, N.; Scarsi, M.; Apostolakis, J.; Ehrhardt, C.; Caflisch, A. *Proteins: Struct., Funct., Bioinformatics* **1999**, *37*, 88−105.
(26) Majeux, N.; Scarsi, M.; Caflisch, A. *Proteins: Struct., Funct., Bioinformatics* **2001**, *42*, 256−268.
(27) Scarsi, M.; Apostolakis, J.; Caflisch, A. *J. Phys. Chem. A* **1997**, *101*, 8098−8106.
(28) Budin, N.; Majeux, N.; Caflisch, A. *Biol. Chem.* **2001**, *382*, 1365−1372.
(29) Cecchini, M.; Kolb, P.; Majeux, N.; Caflisch, A. *J. Comput. Chem.* **2004**, *25*, 412−422.
(30) Kearsley, S. K.; Smith, G. M. *Tetrahedron Comput. Methodol.* **1990**, *3*, 615−633.

Table 1. BACE-1 Inhibitors Identified by High-Throughput Fragment-Based Docking

COMPOUND	STRUCTURE R1	R2	R3	MW (g mol^{-1})	BACE-1[a] IC_{50} (μM)	Abeta(sw)[b] EC_{50} (μM)	Cytotoxic[c] CC_{50} (μM)	LIECE[d] K_i (μM)
								
1	m-H, p-H			466	11.2 ± 0.2	>10	19.1	20.2
2	m-H, p-H			511	11.9 ± 4.9	>10	16.7	74.2
3	m-H, p-H			649	25.5 ± 8.4	9.4	31.2	38.8
4	m-Cl, p-CH_3			598	20.6 ± 1.8	>10	>50	49.4
								
5	F		OH	518	27.9 ± 4.2	16.9 ± 1.8	>50	2.9
6	H		Cl	518	151.8 ± 14.5	18.0 ± 7.6	>50	6.6
7	O-CH3		H	522	66.6 ± 11.0	10.9 ± 5.4	28.0	12.6
8	F		Cl	538	7.1 ± 1.2	>25	>50	12.8

[a] The BACE-1 fluorescence resonance energy transfer assay kit was purchased from PanVera (Madison, WI; catalog no. P2985). BACE-1 activity assays were carried out according to the manufacturer's instructions.Values of average and standard deviation are from three independent experiments. [b] Cell-based assay.[45] [c] Cytotoxic concentration in HEK293 cells (not transgenic).[47] [d] See ref 12.

CHARMM minimization with distance-dependent dielectric function. During minimization, the protein was kept rigid. In the larger of the two screening experiments (306 022 compounds, see Results and Discussion), the minimized poses were re-ranked by the LIECE model using first a spacing of 1.0 Å for the finite-difference Poisson calculation as a filter and finally a grid spacing of 0.3 Å for the top 8000 poses, i.e., 2880 compounds.

Computational Requirements. The LIECE approach requires 26 min (mainly for the finite-difference Poisson calculations with grid spacing of 0.3 Å) of a CPU of a single Opteron 244 (1.8 GHz) for each pose in BACE-1. The total CPU time can be further reduced by first using a coarse grid spacing of 1.0 Å in the finite-difference Poisson calculation, which takes about half a minute. The in silico screening of the 306 022 compound library, i.e., docking and LIECE energy evaluation, took about 10 days on a Beowulf cluster of 100 Opteron 1.8 GHz CPUs.

BACE-1 Enzymatic Assay. The BACE-1 fluorescence resonance energy transfer (FRET) assay was performed as described by the manufacturer (PanVera, P2985) with an incubation time of 30 min. Additional measurements were performed in the presence of detergent or with an incubation time of only 3 min to check for nonspecific effects (e.g., compound aggregation[31,32]). Briefly, fluorescence progress curves of 30 μL reaction volumes were measured on a Tecan GENios reader (Männedorf, Switzerland) upon excitation at 535 nm and emission at 580 nm in 384-well microtiter plates (Corning, 3654). Linear regression analysis was calculated with the Magellan 5.0 software (Tecan Austria GmbH, Salzburg, Austria).

Abeta(sw) (Amyloid *β*40 ELISA) Cell-Based Assay. Swedish APP695 transgenic human embryonic kidney 293 cells (HEK 293) were maintained in Dulbecco's modified Eagle's medium (SIGMA) supplemented with 10% fetal calf serum (Gibco) and 200 μg/mL G418 (Gibco) for continued selection of the stably integrated transgene, as described elsewhere.[5] Briefly, a 400× compound stock solution (dissolved in DMSO) was resuspended in 140 μL of medium lacking G418 and distributed in poly-L-lysine-precoated 96-well cell culture plates (final DMSO concentration 0.25%). Immediately thereafter, 50 000 transgenic HEK 293 cells, resuspended in 20 μL of medium lacking G418, were added to each well. After 2 days of incubation at 37° C and 5% CO_2,

(31) Ryan, A. J.; Gray, N. M.; Lowe, P. N.; Chung, C. W. *J. Med. Chem.* **2003**, *46*, 3448−3451.
(32) McGovern, S. L.; Helfand, B. T.; Feng, B.; Shoichet, B. K. *J. Med. Chem.* **2003**, *46*, 4265−4272.

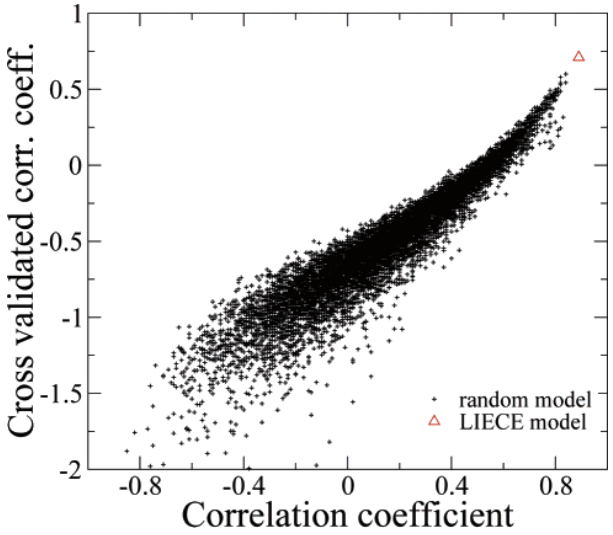


Figure 1. Statistical test to assess the predictive power of the LIECE two-parameter model for BACE-1 compared to 10 000 random models (see main text for details). The fact that the LIECE model data point (red triangle) is on the right-top indicates that LIECE not only better fits the data than the random models (black crosses) but has also a better predictive ability.

an ELISA assay to measure Aβ40 in the supernatant was performed according to the protocol of the manufacturer of the assay kit (The Genetics Company, Switzerland). In parallel, an XTT assay of the cells was performed to measure cell viability, thus verifying that a reduction in the Aβ40 signal is not due to compound toxicity.

Results and Discussion

Validation of the LIECE Model on BACE-1. As in our previous works,[12,33] a two-parameter LIECE model is used here: $\Delta G_{bind} = 0.2737\ \Delta E^{vdW} + 0.1795\ \Delta G^{elec}$, where ΔE^{vdW} is the ligand/protein van der Waals interaction energy and ΔG^{elec} is the sum of the ligand/protein Coulombic energy in vacuo and the change in solvation energy of ligand and protein upon binding. Note that the values of the two parameters ($\alpha = 0.2737$ and $\beta = 0.1795$) were obtained by using a training set of 13 peptidic inhibitors[34] in our previous work[12] and have not been modified since. To further evaluate the predictive power of the LIECE model for BACE-1, three additional tests were performed.

First, a statistical test based on the randomization of the data points was used to analyze an eventual chance correlation.[35,36] The binding free energies of 13 peptidic inhibitors[34] were randomized within the same range as the experimental values, i.e., from -14 to -6 kcal/mol, and the two multiplicative parameters for ΔE^{vdW} and ΔG^{elec} were determined by fitting to random "data points". The randomization and fitting were repeated 10 000 times, and Figure 1 shows the cross-validated correlation coefficient (obtained by the leave-one-out procedure) plotted versus the correlation coefficient. The LIECE model with the two parameters fitted to the real data points is located in the top right corner and is significantly separated from the models generated by the randomization of the binding free energies. This separation provides further evidence that the LIECE two-parameter model not only fits the experimental data but also has very good predictive ability, i.e., chance correlation is not present.

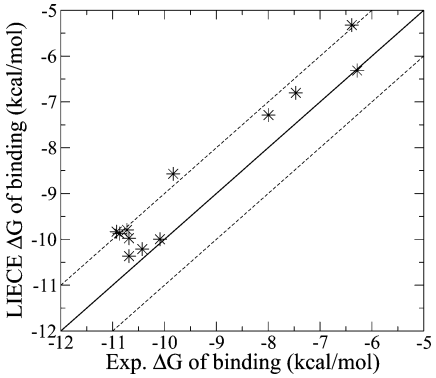


Figure 2. Cross-validation of the LIECE two-parameter model on 12 BACE-1 inhibitors consisting of a 1,3,5-trisubstituted benzene scaffold.[10,38] These 12 BACE-1 inhibitors were not used to derive the LIECE two-parameter model. The dashed lines emphasize the region corresponding to a 1 kcal/mol accuracy.

Second, the recent publication of two X-ray structures of BACE-1 in the complex with nonpeptide inhibitors (PDB codes 1W51[37] and 1TQF[10]) allowed us to perform additional tests of the two-parameter model and its robustness with respect to different protein structures. The LIECE-predicted binding affinity of the 1W51 nonpeptide inhibitor was calculated using two BACE-1 structures, 1W51 and 1M4H.[18] Both calculations gave a LIECE K_i of 0.49 μM, which is close to the experimental IC_{50} of 0.2 μM. Furthermore, we tested a series of 12 inhibitors of BACE-1 based on a 1,3,5-trisubstituted benzene scaffold,[10,38] which adopt a nontraditional binding mode with a displacement of the 10s loop with respect to the 1M4H conformation. These compounds were manually docked into the binding site of 1TQF using the 1TQF inhibitor as a template.[10] The LIECE binding free energy values are plotted versus the corresponding experimental values in Figure 2 (see also Table 1 in the Supporting Information). Remarkably, the root-mean-square of the error and maximal error are 0.78 and 1.3 kcal/mol, respectively, and the correlation coefficient is 0.89. In addition, the LIECE model successfully reproduces the binding energy change between two compounds which differ only in the stereochemistry at the α-methyl group pointing toward the P_3 pocket (compounds **3** and **4** of ref 10). Therefore, the LIECE two-parameter model derived from a single structure (1M4H) shows good predictive ability on a different class of inhibitors, even when the calculations are based on slightly different BACE-1 conformations. This result agrees with the previous work on HIV-1 protease inhibitors binding free energy calculation, where the parameters derived from a single structure were used to accurately predict the activity of a different series of inhibitors.[12] Furthermore, a recent application on 48, 62, and 41 inhibitors of Lck, CDK2, and p38 kinases, respectively, indicates that the LIECE model is transferable among enzymes which share a similar ATP binding site (P. Kolb, D. Huang, F. Dey, and A. Caflisch, manuscript in preparation). Transferability of LIECE parameters between slightly different structures of a given

(33) Huang, D.; Lüthi, U.; Kolb, P.; Edler, K.; Cecchini, M.; Audetat, S.; Barberis, A.; Caflisch, A. *J. Med. Chem.* **2005**, *48*, 5108−5111.
(34) Ghosh, A. K.; Bilcer, G.; Harwood, C.; Kawahama, R.; Shin, D.; Hussain, K. A.; Hong, L.; Loy, J. A.; Nguyen, C.; Koelsch, G.; Ermolieff, J.; Tang, J. *J. Med. Chem.* **2001**, *44*, 2865−2868.
(35) Zoete, V.; Michielin, O.; Karplus, M. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 861−880.
(36) So, S.; Karplus, M. *J. Med. Chem.* **1999**, *39*, 5246−5256.
(37) Patel, S.; Vuillard, L.; Cleasby, A.; Murray, C. W.; Yon, J. *J. Mol. Biol.* **2004**, *343*, 407−416.
(38) Stachel, S. J.; et al. *J. Med. Chem.* **2004**, *47*, 6447−6450.

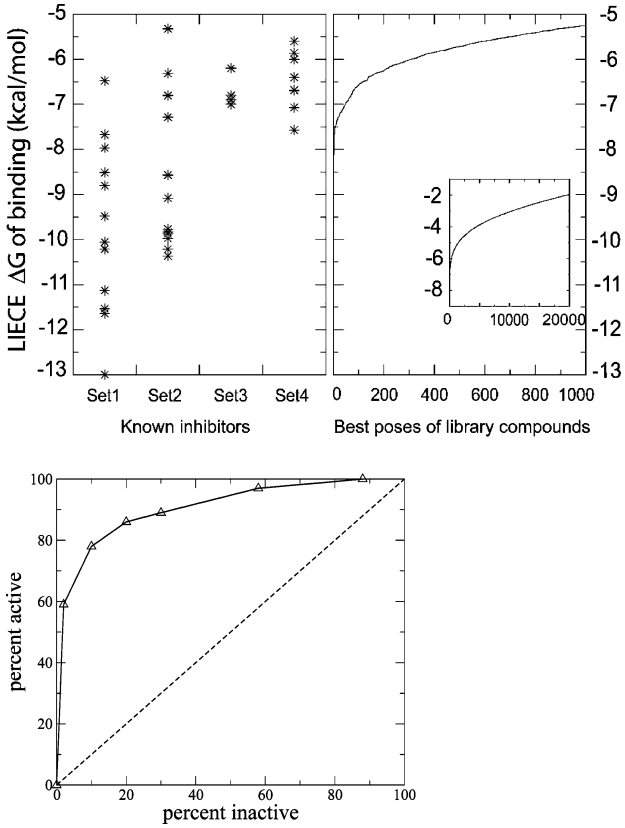


Figure 3. (Top) The LIECE two-parameter model does not generate too many false positives. Comparison between a composite set of 37 known inhibitors of BACE-1 (left) and the 1000 poses with the most favorable LIECE energy from a library of 200 000 small molecules (right), the vast majority of which are not expected to bind to BACE-1. Set1, 13 peptidic inhibitors developed in Tang's group;[34] Set2, 12 derivatives of a 1,3,5-trisubstituted benzene scaffold[10,38] (see also Supporting Information); Set3, four phenylurea derivatives;[33] Set4, eight compounds listed in Table 1. The inset on the right plot zooms out on the first 20 000 poses of the library compounds. (Bottom) Thirty-seven known inhibitors compared to the top 1000 poses as a ROC curve (solid line with triangles). The dashed line of slope 1 shows the behavior of a random model as a basis of comparison. The area under the ROC curve is close to the ideal value, which indicates that the LIECE model generates few false positives.

protein is a useful property, which could be used to take into account binding site flexibility during in silico screening or hit explosion. In this context, we have generated a set of low-energy conformations of BACE-1 using molecular dynamics with explicit water.[39] Because of the transferability of the LIECE parameters, a virtual screening based on these structures may find inhibitors with new binding modes.

Third, it is useful to estimate the amount of false positives, i.e., compounds with good predicted affinity which in reality do not bind. For this purpose, LIECE binding energies of a composite set of 37 BACE-1 inhibitors were compared with those of a library of about 200 000 small molecules (average value of molecular weight of 407.2 ± 72.2 g/mol; this library is unrelated to the libraries used for the in silico screening described here) under the reasonable assumption that very few of the 200 000 compounds inhibit BACE-1. The compounds were docked by FFLD, minimized by CHARMM in the rigid 1M4H structure, and the resulting poses were filtered according to two criteria: the van der Waals intermolecular energy (more favorable than −40 kcal/mol) and the van der Waals intermolecular energy divided by molecular weight (quotient more favorable than −0.1 kcal/g). Figure 3 (top) shows a comparison between the 37 inhibitors (left) and the library of 200 000 compounds (right). Remarkably, 78% and 100% of the known inhibitors have a LIECE energy in the range of values of the 111 and 1000 library compound poses with the most favorable LIECE energy, respectively. The 111 and 1000 poses originate from 100 and 651 different compounds, respectively. In other words, the large majority of the 200 000 compounds are predicted to be worse than most of the known inhibitors. Furthermore, a receiver operating characteristic (ROC) plot[40] for the known 37 compounds over the top 1000 poses (Figure 3 bottom) confirms that the LIECE model of BACE-1 does not generate many false positives.

Effect of Different Protonation States. The protonation state of the catalytic dyad has been investigated by different groups recently.[41,42] The main observation is that only one of the two aspartate side chains should be protonated in the presence of an inhibitor. However, it is still under debate which of the two side chains should be protonated. All calculations in the present study and previous works[12,33] have been performed with Asp228 protonated and Asp32 negatively charged. To test the robustness of this choice, docking of the 306 022 compounds of the second library was repeated using the BACE-1 structure with Asp32 protonated and Asp228 negatively charged. The range of LIECE energies of the top 500 poses (332 compounds) was −8.40 to −5.14 kcal/mol, which is comparable to the previous screening (−8.52 to −5.75 kcal/mol). Importantly, there were 194 compounds in common between the two lists (58%). The four active compounds (**5**−**8**) were ranked in the top 500 list upon docking with Asp32 protonated, and their LIECE affinities were 18.9, 116.8, 105.5, and 54.7 μM, respectively. These values are about an order of magnitude less favorable than those obtained with Asp228 protonated (see below and Table 1).

In Silico Screening and Enzymatic Assay. The DAIM decomposition of the 10 067 and 306 022 compound libraries yielded 469 and 4917 unique fragments, respectively. In the first in silico screening (Figure 4, left), 10 067 compounds were docked, 1000 poses were further evaluated by LIECE energy (1000 unique molecules), 64 compounds (19 of which with a (1,3,5-triazin-2-yl)hydrazone scaffold) were tested in an enzymatic assay, and seven (11%) showed an IC_{50} for BACE-1 smaller than 100 μM. The LIECE ranking of the seven active compounds was among the first 24 of 1000 molecules. Strikingly, the high hit rate was achieved using solely the LIECE energy ranking without manual intervention or visual inspection.

In the second in silico screening (Figure 4, right), 306 022 compounds were docked, 58 000 poses were further evaluated by LIECE (14 085 unique molecules), and 24 compounds (six of which with a (1,3,5-triazin-2-yl)hydrazone scaffold and a carboxy group which was negatively charged for docking and LIECE) were tested in an enzymatic assay. Three of them (12%) showed an IC_{50} smaller than 100 μM, and a fourth compound showed an IC_{50} of 152 μM. Remarkably, these compounds have LIECE ranks of first, fourth, seventh, and eighth.

To obtain information on the mechanism of inhibition and provide evidence against nonspecific effects (e.g., aggrega-

(39) Gorfe, A. A.; Caflisch, A. *Structure* **2005**, *13*, 1487−1498.
(40) Zweig, M. H.; Campbell, G. *Clin. Chem.* **1993**, *39*, 561−577.
(41) Park, H.; Lee, S. *J. Am. Chem. Soc.* **2003**, *125*, 16416−16422.
(42) Rajamani, R.; Reynolds, C. H. *J. Med. Chem.* **2004**, *47*, 5159−5166.

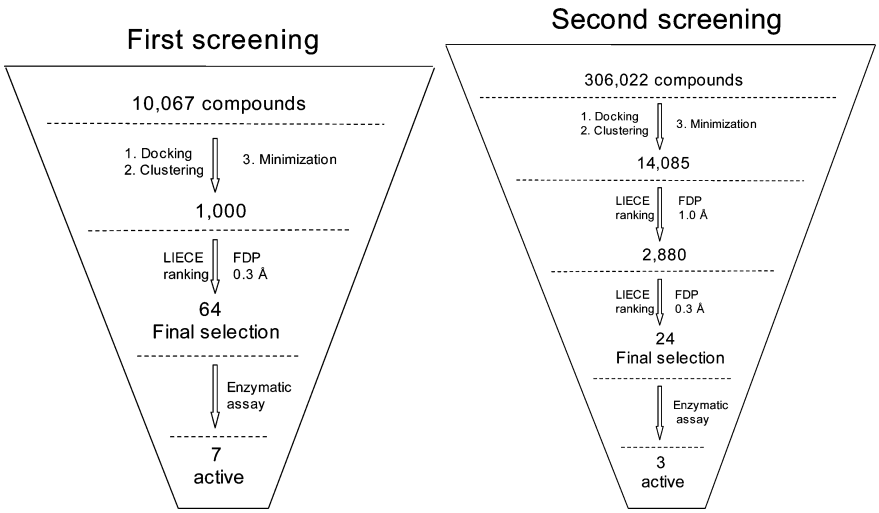


Figure 4. Schematic picture of the two applications of the in silico screening approach. FDP stays for finite-difference Poisson calculations.[17]

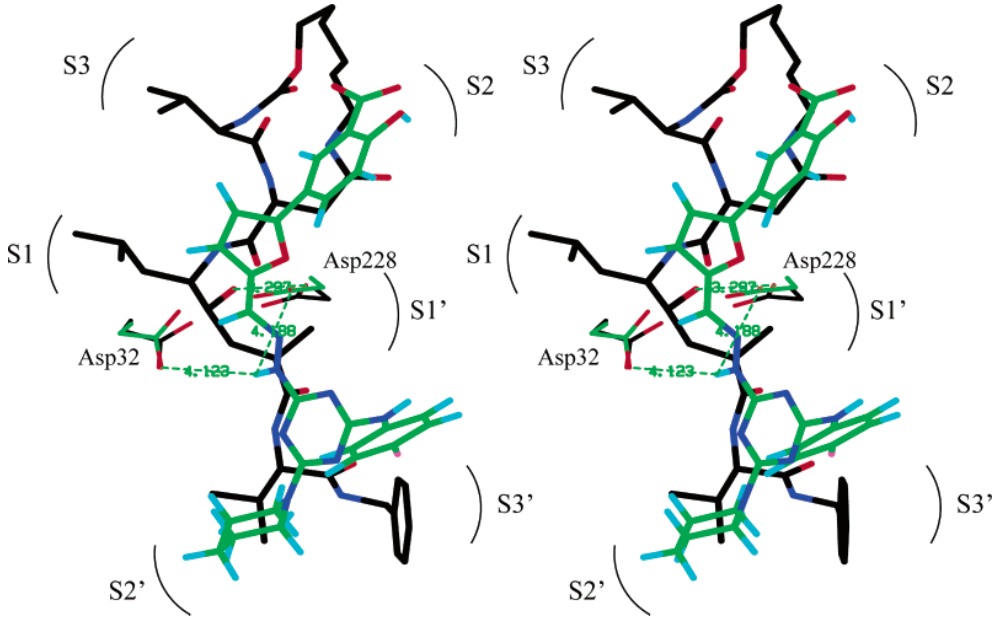


Figure 5. Stereoview of the superposition of a known nanomolar inhibitor of BACE-1[43] (carbon atoms in black) and compound **5** (carbon atoms in green). The side chains of the catalytic residues Asp32 and Asp228 are shown together with the distances between their oxygen atoms and the hydrazone NH of compound **5** or the hydroxyl group of the known nanomolar inhibitor (dashed lines). The structural alignment was generated taking into account only the C_α atoms of the two BACE-1 structures 1XS7[43] and 1M4H,[18] which overlap with a deviation of only 0.3 Å. The 1M4H structure was used for the high-throughput docking.

tion[31,32] or covalent modification), two additional experiments were performed. First, detergent (0.05% (v/v) Triton X-100) was added in the enzymatic assay. No significant reduction of activity was observed. Second, the effect of two different incubation times, 3 vs 30 min, was investigated for compounds **1** and **5**. After the shorter incubation time, the percentage inhibition at 20 μM concentration of compound **1** is 57%, and that at 50 μM concentration of compound **5** is 66%. These values are consistent with the IC_{50} values measured with 30 min incubation time (Table 1). Furthermore, two compounds were tested using another commercially available FRET assay kit (SIGMA CS0010, which includes Triton X-100 0.08% at final concentration). With the SIGMA kit, IC_{50} values of 7 and 32 μM were measured for compounds **7** and **8**, respectively. Compound **7** shows 1 order of magnitude difference in the IC_{50} value measured with two different kits. This discrepancy is likely to be a consequence of differences in substrate, protein, and assay buffer.

Interestingly, the 11 active compounds (10 with $IC_{50} < 100$ μM and one with $IC_{50} = 152$ μM) from the two in silico screenings have a common (1,3,5-triazin-2-yl)hydrazone scaffold. Table 1 shows structure as well as experimental and predicted affinity of eight compounds, four from each screening. Compounds **1**−**3** differ only in the substituents of the ring at R3, and compounds **5** and **6** are also very similar.

Binding Mode. The predicted binding mode of compound **5** is shown in Figure 5, overlapped with the cycloamide-urethane-derived peptidic inhibitor **2c** of ref 43. The hydrogen atom of the hydrazone NH group is at a distance of about 4 Å from two oxygen atoms in the catalytic aspartates. Such distance suggests

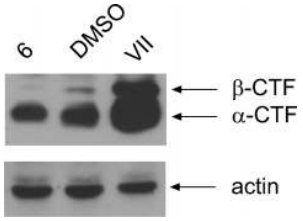


Figure 6. Western blot analysis of compound **6**, confirming reduction of the β-site cleavage of APP. (Top) The β-carboxy-terminal fragment (CTF) and α-CTF bands indicate that cleavage of β-CTF is decreased compared to DMSO negative control without affecting α-CTF. On the other hand, the γ-secretase inhibitor VII[46] causes an accumulation of both β-CTF and α-CTF. (Bottom) The actin bands indicate equal loading of the samples in each lane.

the presence of a water-bridged hydrogen bond as observed in the X-ray structure of BACE-1 in the complex with an oxyacetamide compound (IC_{50} = 1.4 μM, PDB code 1TQF[10]) and in the structure of plasmepsin II complexed with an inhibitor featuring a tertiary amino group close to the two catalytic aspartates.[44] The S2, S2′, and S3′ pockets of BACE-1 are occupied by the 2-hydroxybenzoic acid, piperidine, and fluorobenzene substituents of compound **5**, respectively. The 2-hydroxybenzoic acid and piperidine of **5** overlap with part of the macrocycle and P2′-propyl side chain of **2c**, respectively. On the other hand, the fluorobenzene of **5** has a slightly different orientation compared to the benzyl group of **2c**. The furan and (1,3,5-triazin-2-yl)hydrazone mimic part of the peptidic backbone of **2c**.

The predicted binding mode of compounds **6**−**8** is essentially identical to the one of **5**, while the R3 substituent of compounds **1**−**4**, which lack the furan linker, points toward the S1 pocket. Given the small range of measured IC_{50} values and uncertainties in the details of the predicted binding mode (determined by FFLD docking and energy minimization in the rigid BACE-1 structure), it is not possible to obtain a detailed structural explanation of the measured relative affinities. The rather small differences in measured affinities is consistent with the fact that compounds **1**−**8** are similar among each other and show similar predicted binding modes.

Cellular Assay. To assess the potential for further development, e.g., hit explosion, it is important to verify that the compounds which are active in the enzymatic assay are also cell-permeable and able to inhibit BACE-1 in mammalian cells. For this purpose, 7 and 24 compounds from the first and second screening, respectively, were submitted to a cell-based test in which reduction of Aβ peptide secretion was measured as reported previously by others.[45] Table 1 shows that compounds **3** and **5**−**7** from the first and second screenings are active, with EC_{50} < 10 μM and EC_{50} < 20 μM, respectively.

It is interesting to note that compound **8**, with the highest potency in the enzymatic assay for BACE-1 (IC_{50} = 7.1 μM in the Panvera kit and 32 μM in the SIGMA kit), is not active in the cell-based assay at a concentration of 25 μM. These data have to be compared with the corresponding ones for compound **6**, which has a very poor activity in the enzymatic assay but a cell-based EC_{50} value of 18.0 μM. These discrepancies might be due to several reasons, including differences in cell permeability, cleavage efficiency of full-length BACE-1 (cell-based assay) versus the lumenal domain only (enzymatic assay), different substrates, and assay conditions (e.g., pH 4.5 in the enzymatic assay).

To provide further evidence that compound **6** inhibits BACE-1 activity in cells, Western blot analysis was used to detect differentially cleaved carboxy-terminal fragments (CTFs) of APP (Figure 6). Compound **6** lowered β-CTF without affecting α-CTF compared to the dimethyl sulfoxide (DMSO) negative control. On the other hand, the γ-secretase inhibitor VII[46] caused an accumulation of both β-CTF and α-CTF compared to DMSO. This result indicates that reduced secretion of Aβ from cells was caused by β-site cleavage inhibition.

Finally, it is important to note that the peptidic inhibitors OM99-2 (molecular weight 897.2 g/mol),[34] its cycloamide-urethane derivative **2c** (731.1 g/mol),[43] and the peptidomimetic **57** (699.4 g/mol) of ref 11 have low-nanomolar affinity for BACE-1 in enzymatic tests but show only micromolar activity in cellular assays because of limited ability to cross cell membranes.[11,43] Despite their more than 3 orders of magnitude worse inhibitory activity in the enzymatic assay, the four triazine derivatives **3** and **5**−**7** have cellular activity similar to that of the three known peptidic inhibitors mentioned above. Given their smaller size, the triazine derivatives are likely to be more suitable for further development than the peptidic inhibitors.

Conclusions

High-throughput, fragment-based docking into the BACE-1 active site and LIECE binding free energy evaluation were used to select 88 compounds for experimental validation from an initial set of more than 300 000 molecules. Ten of the 88 compounds inhibit BACE-1 activity in an enzymatic assay (IC_{50} < 100 μM), and four of them are active in a mammalian cell-based assay (EC_{50} < 20 μM). Taken together, the discoveries of three novel series of BACE-1 inhibitors, i.e., phenylurea derivatives,[33] triazine derivatives (this work), and a set of five cell-permeable, nonpeptide, low-micromolar inhibitors of BACE-1 with a different scaffold (D. Huang and A. Caflisch, unpublished results), are a proof-of-principle of our in silico high-throughput screening approach. Furthermore, the present study represents a successful combination of computational predictions and experimental validation of inhibitors of a pharmaceutically relevant enzyme for which few nonpeptidic inhibitors have been already discovered, despite the availability of the X-ray structure of BACE-1 for more than 5 years. We are currently applying high-throughput docking and the LIECE approach to identify kinase inhibitors from very large collections of low-molecular-weight compounds. For protein targets of known three-dimensional structure, the efficient in silico approach presented in this paper is a cost-effective alternative to high-throughput in vitro screening campaigns.

Availability of the Software. The software suite of programs for high-throughput docking (DAIM, SEED, FFLD), including

(43) Ghosh, A.; Devasamudram, T.; Hong, L.; DeZutter, C.; Xu, X.; Weerasena, V.; Koelsch, G.; Bilcer, G.; Tang, J. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 15−20.
(44) Prade, L.; Jones, A. F.; Boss, C.; Richard-Bildstein, S.; Meyer, S.; Binkert, C.; Bur, D. *J. Biol. Chem.* **2005**, *280*, 23837−23843.
(45) Dovey, H. R.; Suomensaari-Chrysler, S.; Lieberburg, L.; Sinha, S.; Keim, P. S. *NeuroReport* **1993**, *4*, 1039−1042.
(46) Durkin, J. T.; Murthy, S.; Husten, E. J.; Trusko, S. P.; Savage, M. J.; Rotella, D. P.; Greenberg, B. D.; Siman, R. *J. Biol. Chem.* **1999**, *274*, 20499−20504.
(47) Scudiero, D. A.; Shoemaker, R. H.; Paull, K. D.; Monks, A.; Tierney, S.; Nofziger, T. H.; Currens, M. J.; Seni, D.; Boyd, M. R. *Cancer Res.* **1988**, *48*, 4827−4833.

input files, test cases, and documentation, are available from the corresponding author (at no expense for not-for-profit institutions).

Acknowledgment. We are grateful to Stephan Audetat, Fabian Dey, and Nicolas Majeux for interesting discussions and Karin Edler for help with the in vitro experiments. The calculations were performed on Matterhorn, a Beowulf Linux cluster at the Informatikdienste of the University of Zurich, and we thank C. Bolliger, T. Steenbock, and A. Godknecht for installing and maintaining the Linux cluster. This work was supported by grants from KTI (Kommission Technologie und Innovation) and the National Competence Center for Research (NCCR) on Neural Plasticity and Repair.

Supporting Information Available: Experimental procedures, information on 88 tested compounds, the LIECE energy values of 37 known inhibitors and the top 100 library compounds of Figure 3, and complete refs 3, 11, and 38. This material is available free of charge via the Internet at http://pubs.acs.org.

JA0573108

Chapter 6

Replica Exchange Molecular Dynamics Simulations of Amyloid Peptide Aggregation

(Journal of Chemical Physics 121, pp 10748-10756, 2004)

Replica exchange molecular dynamics simulations of amyloid peptide aggregation

M. Cecchini, F. Rao, M. Seeber, and A. Caflisch[a)]
Department of Biochemistry, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland

(Received 6 May 2004; accepted 1 September 2004)

The replica exchange molecular dynamics (REMD) approach is applied to four oligomeric peptide systems. At physiologically relevant temperature values REMD samples conformation space and aggregation transitions more efficiently than constant temperature molecular dynamics (CTMD). During the aggregation process the energetic and structural properties are essentially the same in REMD and CTMD. A condensation stage toward disordered aggregates precedes the β-sheet formation. Two order parameters, borrowed from anisotropic fluid analysis, are used to monitor the aggregation process. The order parameters do not depend on the peptide sequence and length and therefore allow to compare the amyloidogenic propensity of different peptides. © *2004 American Institute of Physics.* [DOI: 10.1063/1.1809588]

I. INTRODUCTION

A thorough sampling of conformational space is required to describe the thermodynamics of complex systems such as multiple peptide chains at finite concentrations. Constant temperature molecular dynamics (CTMD) techniques often fail to adequately sample conformational space of frustrated and minimally frustrated systems which are characterized by a rugged free-energy landscape where energy barriers between minima are higher than the thermal energy at physiological temperature. For this reason, a number of approaches to enhance sampling of phase space have been introduced.[1–4] The parallel tempering technique (also known as replica exchange) was developed for dealing with the slow dynamics of disordered spin systems.[5] Sugita and Okamoto have extended the original formulation of replica exchange into an MD based version (REMD) and tested it on the pentapeptide Met-enkephalin *in vacuo*.[6] Although in the context of fragile liquids De Michele and Sciortino found that parallel tempering does not increase the speed of equilibration of the (slow) configurational degrees of freedom,[7] in the case of atomistic simulations of proteins many different applications have shown the efficiency of the method. Sanbonmatsu and Garcia have used REMD to investigate the structure of Met-enkephalin in explicit water,[8] and the α-helical stabilization by the arginine side-chain which was found to originate from the shielding of main chain hydrogen bonds.[9] REMD has also been applied to investigate the energy landscape of the C-terminal β-hairpin of protein G[10,11] and a three-helix bundle protein.[12] REMD in implicit solvent has been used to investigate the thermodynamics of designed 20-residue structured peptides,[13,14] and recently to study folding of a helical transmembrane protein.[15]

Highly ordered protein aggregates are associated with severe human disorders including Alzheimer's disease, type-II diabetes, systemic amyloidosis, and transmissible spongiform encephalopathies.[16,17] The soluble precursors of the ordered protein deposits do not share any sequence homology or common fold. However, x-ray diffraction data indicate a cross-β-structure for most fibrillar aggregates.[18,19] These findings suggest that key steps in the aggregation process may be common to all amyloidogenic proteins. Despite the medical relevance of amyloidoses, many important questions about the formation of ordered aggregates remain unanswered. There is experimental evidence that cytotoxicity is more pronounced for the early aggregates than for highly organized fibrillar structures.[20] Moreover, some peptide fragments of amyloidogenic proteins display the same properties as the full-length protein, including cooperative kinetics of aggregation, fibril formation, binding of the dye Congo red, and the cross-β x-ray diffraction pattern.[21] Both findings are particularly interesting because current simulation approaches allow significant sampling only for oligomeric peptide systems.

There have been several lattice studies on aggregation in proteins. These simplified models have allowed to investigate the foldability and aggregation propensity[22,23] and how interaction potentials affect the properties of aggregation-prone proteins.[24] Harrison *et al.* have shown that less stable proteins have a greater chance of assuming alternative native states as multimers.[25] MD simulations of aggregation have been performed by using a three-bead backbone and single-bead side-chain model.[26] While this simplified model has allowed the simulation of the competition between folding and aggregation for two four-helix bundles it is probably not possible to extract detailed information on energetics and sequence dependence. Recently, a minimalist Go model of four peptide strands[27] has been investigated by MD simulations in a confining sphere and the aggregation process was shown to depend on both sequence and environment.[28] Atomic models of amyloidogenic peptides have been simulated by MD with an implicit treatment of the solvent[29–31] and explicit water molecules.[32–36]

Recently, a replica exchange Monte Carlo technique has

[a)]Author to whom correspondence should be addressed. Fax: +41 1 635 68 62; Electronic mail: caflisch@bioc.unizh.ch

been applied to a lattice Go model of a minimalist multichain system to study the interplay between folding and disordered aggregation[23] but atomic model REMD applications to ordered aggregation have not been reported yet.

In the present paper, REMD with implicit solvent[37] is used to investigate the thermodynamics of the early steps of peptide aggregation and comparison is made with CTMD. The present work was motivated by three questions. Is it possible to sample the early events of ordered peptide aggregation at physiologically relevant temperatures? Do the aggregation energetics sampled by REMD correspond to those observed in CTMD simulations? Are the nematic and polar order parameters, borrowed from liquid crystal theory, useful to describe aggregation? The simulation results indicate that all questions can be answered affirmatively. Moreover, the ''liquid crystal'' order parameters allow to discriminate amyloidogenic peptide sequences from those that form only disordered aggregates.

II. METHODS

A. Model

The MD simulations and part of the analysis of the trajectories were performed with the CHARMM program.[38] The oligomeric peptide systems were modeled by explicitly considering all heavy atoms and the hydrogen atoms bound to nitrogen or oxygen atoms (PARAM19 potential function[38,39]). The remaining hydrogen atoms are considered as part of the carbon atoms to which they are covalently bound (extended atom approximation). The effective energy, whose negative gradient corresponds to the force used in the dynamics, is

$$E(\mathbf{r}) = E_{vacuo}(\mathbf{r}) + G_{solv}(\mathbf{r}) \tag{1}$$

for a molecular system with atomic nuclei located at $\mathbf{r}=(\mathbf{r}_1,\ldots,\mathbf{r}_N)$. The PARAM19 *vacuo* energy function is

$$\begin{aligned} E_{vacuo}(\mathbf{r}) = {} & \frac{1}{2}\sum_{bonds} k_b(b-b_0)^2 + \frac{1}{2}\sum_{\substack{bond\\ angles}} k_\theta(\theta-\theta_0)^2 \\ & + \frac{1}{2}\sum_{\substack{dihedral\\ angles}} k_\phi[1+\cos(n\phi-\delta)] \\ & + \frac{1}{2}\sum_{\substack{improper\\ dihedrals}} k_\omega(\omega-\omega_0)^2 \\ & + \sum_{i>j} \varepsilon_{ij}^{\min}\left[\left(\frac{d_{ij}^{\min}}{r_{ij}}\right)^{12} - 2\left(\frac{d_{ij}^{\min}}{r_{ij}}\right)^{6}\right] \\ & + \sum_{i>j} \frac{q_i q_j}{\epsilon(r_{ij}) r_{ij}}, \end{aligned}$$

where b is a bond length, θ a bond angle, ϕ a dihedral angle, ω an improper dihedral, r_{ij} is the distance between atoms i and j, q_i and q_j are partial charges, and $d_{ij}^{\min}$ and $\varepsilon_{ij}^{\min}$ are the optimal van der Waals distance and energy, respectively. Parameters are given in Ref. 39.

An implicit model based on the solvent accessible surface was used to describe the main effects of the aqueous solvent on the solute.[37] In this approximation, the solvation free energy is given by

$$G_{solv}(\mathbf{r}) = \sum_{i=1}^{N} \sigma_i A_i(\mathbf{r}) \tag{2}$$

for a molecular system having N heavy atoms with Cartesian coordinates $\mathbf{r}=(\mathbf{r}_1,\ldots,\mathbf{r}_N)$. $A_i(\mathbf{r})$ is the solvent-accessible surface computed by an approximate analytical expression[40] and using a 1.4 Å probe radius. The solvation model contains only two σ parameters: one for carbon and sulfur atoms ($\sigma_{C,S}=0.012$ kcal/mol Å^2), and one for nitrogen and oxygen atoms ($\sigma_{N,O}=-0.060$ kcal/mol Å^2).[37] Hence, according to Eq. (2) hydrophobic side chains tend to be buried within the solute whereas hydrophilic side chains and the polar groups of the backbone prefer to be solvent accessible. Furthermore, ionic side chains were neutralized[41] and a linear distance-dependent screening function [$\epsilon(r_{ij})=2r_{ij}$] was used for the electrostatic interactions. The CHARMM PARAM19 default cutoffs for long range interactions were used, i.e., a shift function[38] was employed with a cutoff at 7.5 Å for both the electrostatic and van der Waals terms. This cutoff length was chosen to be consistent with the parametrization of the force-field and implicit solvation model. The model is not biased toward any particular secondary structure type. In fact, exactly the same force field and implicit solvent model have been used recently in MD simulations of aggregation,[30,31] folding of structured peptides (α-helices and β-sheets) ranging in size from 15 to 31 residues,[42–44] and small proteins of about 60 residues.[45,46]

B. REMD simulations

The basic idea of REMD is to simulate different copies (*replicas*) of the system at the same time but at different temperatures values. Each replica evolves independently by MD and every t_{swap} states i,j with neighbor temperatures are swapped (by velocity rescaling) with a probability $w_{ij}=\exp(-\Delta)$,[6] where $\Delta\equiv(\beta_i-\beta_j)(E_j-E_i)$, $\beta=1/kT$, and E is the effective energy [potential and solvation energy, Eq. (1)]. A t_{swap} of 10 000 MD steps (20 ps) was chosen in order to allow the kinetic and potential energy of the system to relax. High temperature simulation segments facilitate the crossing of the energy barriers while the low temperature ones explore in detail energy minima. The result of this swapping between different temperatures is that high temperature replicas help the low temperature ones to jump across the energy barriers of the system.

In this study six replicas were used with temperatures (in kelvin) 275, 296, 319, 344, 371, and 400. This range corresponds to a subset of values used in a previous study of reversible peptide folding with the same force-field and solvation model.[14] The acceptance ratios of exchange between neighbor temperatures ranged between 15% and 24%. Each trajectory has a length of 2 μs for a total of 12 μs of simulation time (see Table I).

TABLE I. Simulations performed.

Peptide sequence	Length (μs)	T (K)	Method	IP aggregation events	IA aggregation events
GNNQQNY	10×0.5	275	CTMD	0	6 (19.2)[a]
GNNQQNY	5×1.0	296	CTMD	3 (14.4)	5 (1.6)
GNNQQNY	10×3.4	330	CTMD	54 (7.6)	43 (1.4)
GNNQQNY	2×1.0	371	CTMD	0	0
GNNQQNY	6×2.0	275–400	REMD	14 (60.3)	15 (3.9)
QQQQQQQ	6×2.0	275–400	REMD	27 (54.8)	2 (9.4)
AAAAAAA	6×1.0	275–400	REMD	4 (0.8)	12 (0.9)
SQNGNQQRG	6×2.0	275–400	REMD	1 (1.6)	6 (1.0)

[a]The average time (ns) the three peptides remained aggregated in IP and IA is given in parentheses.

C. Constant temperature MD simulations

A series of control runs were performed at constant temperature: (i) ten simulations at 330 K (total of 34 μs) used as a comparison for the aggregation process between CTMD and REMD (see Table I), (ii) ten 0.5 μs simulations at 275 K and (iii) five 1 μs simulations at 296 K to compare CTMD and REMD sampling at physiologically relevant conditions, and (iv) two 1 μs simulations at 371 K to study the system near the *condensation* temperature (see below).

For both REMD and CTMD, Langevin dynamics with a friction value of 0.15 ps^{-1} was used. This friction coefficient is much smaller than the one of water (43 ps^{-1} at 330 K computed as $3\pi\eta d/m$,[47] where η is the viscosity of water at 330 K, and d and m are the effective diameter, i.e., 2.8 Å, and mass of a water molecule, respectively) to allow for sufficient sampling within the μs time scale of the simulation. The small friction does not influence the thermodynamic properties of the system.

The SHAKE algorithm[48] was used to fix the length of the covalent bonds involving hydrogen atoms, which allows an integration time step of 2 fs. Furthermore, the nonbonded interactions were updated every ten dynamics steps and coordinate frames were saved every 20 ps for a total of 5×10^4 conformations/μs. A 1 μs run requires approximately two weeks on a 1.4 GHz Athlon processor and the REMD simulations were run in parallel on a Linux Beowulf cluster.

D. Progress variables

Aggregation contacts. In-register parallel and antiparallel aggregation contacts were defined following the prescription given in Ref. 30: a contact was considered to be present if the distance between two C_α atoms placed on different in-register strands was within 5.5 Å. The fraction of in-register parallel contacts Q_p and in-register antiparallel contacts Q_a were used to monitor the evolution of the aggregation process. In-register parallel and antiparallel aggregates, IP and IA, respectively, were considered formed when Q_p and Q_a were larger than 0.75 ($Q_p, Q_a > 11/14$) whereas at values smaller than 0.25 ($Q_p, Q_a < 4/14$), the system was considered disordered. The aggregation time is defined as the temporal interval between the first time point where $Q_p, Q_a < 0.25$ and the following time point where $Q_p, Q_a > 0.75$.

Radius of gyration. The radius of gyration of the oligomeric system R_g was considered to monitor the degree of *condensation* and calculated using the minimum image convention. Large values of R_g indicate conformations with isolated and non-interacting peptides (*uncondensed phase*). Small values of R_g indicate ordered as well as disordered aggregated conformations (*condensed phase*).

E. Orientational order parameters

The nematic and polar order parameters, $\overline{P_2}$ and $\overline{P_1}$, respectively, were considered in this study. These order parameters represent the first and second rank coefficients of the singlet orientational distribution expanded in a Wigner series,[49,50] i.e., a basis set of the Wigner rotation matrices. The nematic and polar order parameters are widely used for studying the properties of anisotropic fluids such as liquid crystals[51–54] and are defined as

$$\overline{P_2} = \frac{1}{N}\sum_{i=1}^{N} \frac{3}{2}(\hat{\mathbf{z}}_i \cdot \hat{\mathbf{d}})^2 - \frac{1}{2} \tag{3}$$

and

$$\overline{P_1} = \frac{1}{N}\sum_{i=1}^{N} \hat{\mathbf{z}}_i \cdot \hat{\mathbf{d}}, \tag{4}$$

where $\hat{\mathbf{d}}$ (the director) is a unit vector defining the preferred direction of alignment, $\hat{\mathbf{z}}_i$ is a suitably defined molecular vector, and N is the number of molecules in the simulation box, i.e., three peptides in this study. The director is defined as the eigenvector of the ordering matrix,[55] that corresponds to the largest eigenvalue. Here, the molecular vectors $\hat{\mathbf{z}}_i$ were defined as unit vectors linking the peptide's termini (from the N to the C terminus, Fig. 1). To optimally select the $\hat{\mathbf{z}}_i$ vectors, other choices were investigated: vectors linking the carbonyl C to the amide N of each residue (''amide'' vectors) as well as vectors lying along the carbonyl bonds. Similar results were obtained with the three different choices of $\hat{\mathbf{z}}_i$. However, due to the atomic connectivity along the backbone the amide vectors are not fully independent; their orientations are strongly correlated and the description of the ordered macrostates results less precise. The same is true for the ''carbonyl'' vectors. Hence, vectors linking peptide's termini were preferred.

The order parameters [Eqs. (3) and (4)] change value on going from one order macrostate to the other and should vanish when the transition to a fully isotropic state takes place. They describe different orientational properties of the system and yield useful and complementary information. The

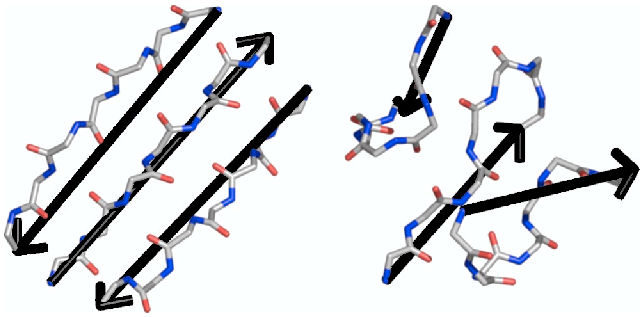


FIG. 1. Pictorial representation of the molecular vectors $\hat{\mathbf{z}}_i$ (black arrows) used to compute the order parameters $\overline{P_1}$ and $\overline{P_2}$. $\hat{\mathbf{z}}_i$ vectors are defined as full length peptide vectors (linking the peptide's termini) and allow to clearly discriminate between ordered (left, $\overline{P_2}=0.87$) and disordered (right, $\overline{P_2}=0.46$) conformations of the system. [The pictures were drawn using the program PYMOL (Ref. 66)].

nematic $\overline{P_2}$ describes the orientational order of the system and discriminates between ordered and disordered conformations. The polar $\overline{P_1}$ describes the polarity of the system, i.e., how much the molecular vectors $\hat{\mathbf{z}}_i$ point in the same direction, and discriminates between parallel and antiparallel/mixed ordered aggregates.

F. Peptides

To evaluate the reliability of amyloidogenic propensity estimations, four oligomeric peptide systems were considered in this study: the amyloid-forming heptapeptide GNNQQNY and the soluble nonapeptide SQNGNQQRG both from the yeast prion Sup35 (residues 7–13 and 17–25 with the Gln/Arg mutation at position 24, respectively),[21] the amyloidogenic poly(L-glutamine) QQQQQQQ (Ref. 56) and the nonamyloidogenic poly(L-alanine) AAAAAAA.[57] To reproduce the experimental conditions,[21,56,57] the peptide systems derived from the yeast prion Sup35 were modeled without blocking groups, while the Ala and Gln repeats were both *N*-acetylated and C-amidated.

All simulations were performed with three peptide replicas starting from random conformations, positions, and orientations. In the initial random positions there was no intermolecular contact, i.e., the peptides were separated in space. Each system was simulated in a cubic box of 75 Å per side yielding a sample concentration of 0.012 M. Since the oligomeric systems present different molecular weights, the above reported concentration corresponds to 3.4, 3.9, 5.4, and 3.4 mg/ml for GNNQQNY, SQNGNQQRG, QQQQQQQ, and AAAAAAA, respectively.

G. Analysis tools

The aggregation contacts, radius of gyration, and order parameters analysis was carried out with a GPL licensed program[58] developed in house to manipulate and analyze molecular dynamics (MD) trajectories. The program is optimized for speed and ease of usage so that it allows extensive processing of large amounts of data and straightforward addition of new analysis tools. Compared to other available programs,[38,59] the analysis of MD trajectories is much faster.

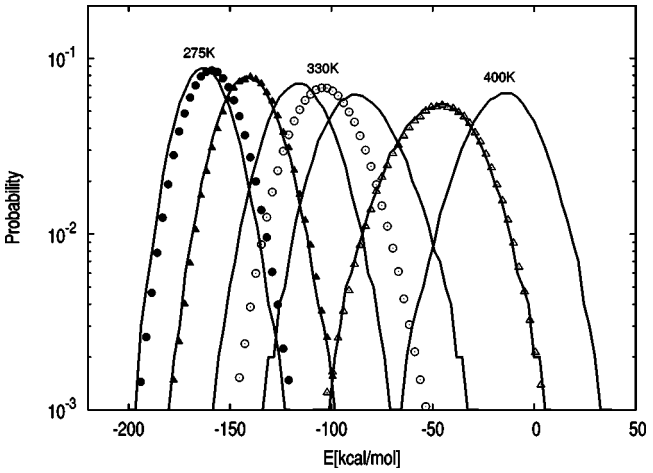


FIG. 2. Probability distribution of the effective energy for the REMD (solid lines) and the CTMD control simulations (filled circles, filled triangles, empty circles, and triangles for 275, 296, 330, and 371 K, respectively). The REMD distributions correspond to the following temperatures (from left to right): 275, 296, 319, 344, 371, and 400 K. The asymmetry of the curves and the temperature dependence of the distributions indicate the presence of a phase transition around 371 K (see text).

III. RESULTS AND DISCUSSION

A. REMD diagnostics

The set of temperatures used in a REMD simulation is crucial for a correct and efficient sampling.[8] Since a simple *a priori* protocol for selecting the optimal temperature distribution has not been identified (yet), the choice often follows empirical considerations:[8,14,23] the highest temperature of the set has to be high enough to overcome energy barriers, while the lowest temperature has to allow the exploration of minima. However, given a fixed number of replicas the temperature range cannot be too wide. Temperature values need to be close enough to make the energy histograms overlap (see Fig. 2) in order to guarantee a high number of temperature swaps during a simulation run. In this study, a set of six temperature values ranging from 275 to 400 K has been selected (see Methods). The time series of temperature exchanges for one of the six replicas is shown in Fig. 3. During the simulation, each replica visits all the temperatures of the set several times realizing the desired free random walk in temperature space.[6]

Symbols in Fig. 2 show the results from CTMD simulations carried out at 275 (filled circles), 296 (filled triangles), 330 (empty circles), and 371 K (empty triangles). At 330 K, the CTMD effective energy distribution is located between the REMD distributions extracted at 319 and 344 K and shows a consistent functional profile. At 371 K, CTMD and REMD effective energy distributions overlap. Therefore, the energetic properties of an aggregating system sampled by a REMD simulation at medium and high temperatures correspond to those observed in CTMD simulations. However, approaching the physiologically relevant conditions the CTMD distributions tend to shift toward less favorable energies (Fig. 2, filled symbols). CTMD at low temperature can get trapped in local energy minima and REMD is superior in sampling conformational space.[6,14]

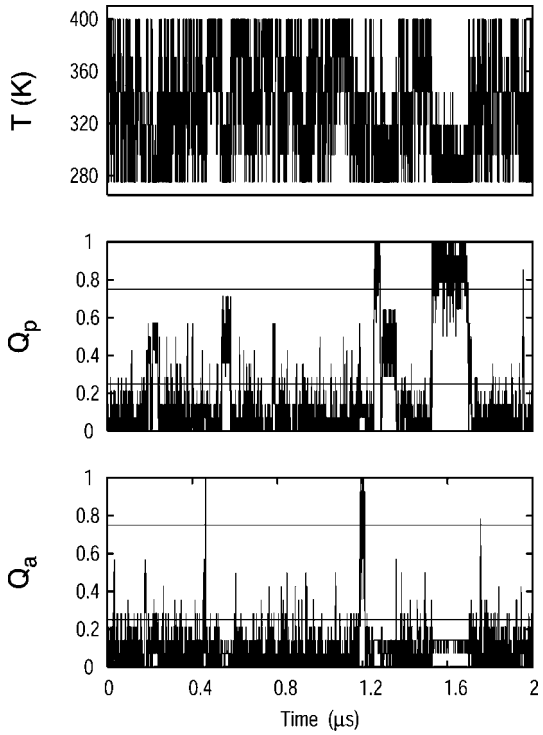


FIG. 3. Time series of (from top to bottom) the temperature T, the fraction of in-register parallel contacts Q_p, and the fraction of in-register antiparallel contacts Q_a for a REMD replica. Along the trajectory, replicas realize the desired free random walk in temperature space (top) so that an efficient sampling of the ordered aggregates is allowed (peaks in Q_p and Q_a plots). Horizontal lines in the time series of the fraction of aggregation contacts indicate the upper/lower thresholds used to define the ordered aggregation/disaggregation events.

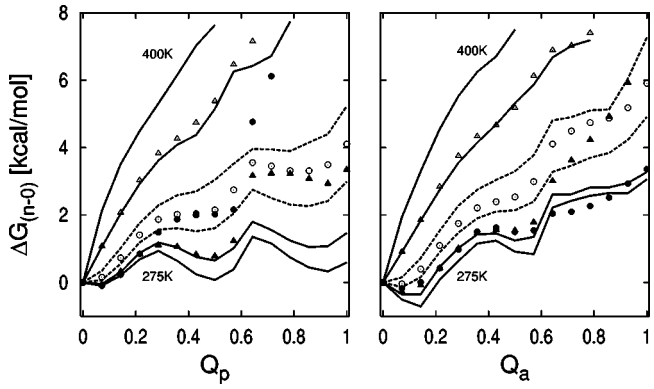


FIG. 4. Free-energy projections along the fraction of in-register parallel contacts Q_p (left) and in-register antiparallel contacts Q_a (right). Conformations with zero in-register contacts were chosen as reference states. $\Delta G_{(n-0)}$ was computed as $-k_B T \ln(N_n/N_0)$, where N_n indicates the number of conformations with n contacts and k_B is the Boltzmann constant. REMD data are shown in solid lines for all the temperature values except for 319 and 344 K which are in dashed lines. CTMD data are shown with symbols (filled circles, filled triangles, empty circles, and triangles for 275, 296, 330, and 371 K, respectively).

The time series of the fraction of in-register parallel contacts (Q_p) and in-register antiparallel contacts (Q_a) have been monitored along the REMD trajectories (Fig. 3). A total of 14 IP and 15 IA aggregation events have been observed along the total simulation time of 12 μs (see Table I). The average aggregation time (see Methods) was 0.74 μs for IP and 0.75 μs for IA arrangements. The average aggregation time determined from the REMD simulation is similar to the values obtained from 34 μs CTMD simulations at 330 K. It is worth noting that in a preliminary REMD run with higher temperatures values (6×1 μs, 319–465 K; data not shown) only 3 IP and 4 IA aggregation events were sampled. The temperature range is crucial in REMD and it has to be carefully chosen in order to speed up the conformational search of relevant states,[60] i.e., the *ordered states* when studying aggregation. To bias the search toward conditions where ordered states are more probable, the temperature was set to lower values (275–400 K, as mentioned above) and the sampling of aggregation events turned out substantially improved.

Figure 4 shows the projections of the free energy surface along Q_p and Q_a for both REMD and CTMD trajectories. The profiles indicate that the structural properties of the aggregating system sampled by a REMD simulation correspond to those observed in CTMD simulations only at high and medium temperatures. At 371 K, CTMD and REMD free energy projections overlap. At 330 K, the CTMD free-energy profiles (empty circles) are correctly placed between REMD projections at 319 and 344 K (dashed lines) and show patterns characterized by a well-defined local minimum at $Q_p > 0.7$ and a monotonic uphill trend along Q_a, fully consistent with the profiles extracted from the REMD simulation. However, at low temperature (275 and 296 K) the free energy profiles extracted from CTMD and REMD trajectories are not consistent any more and the most ''relevant'' conformations, which correspond to in-register parallel and antiparallel arrangements ($Q_p, Q_a > 0.7$), are not correctly sampled by CTMD (Fig. 4, filled symbols).

B. Temperature dependence of ordered amyloid peptide aggregation

Since the energetic and structural properties of the system are not artificially altered (see preceding section), the REMD approach allows to evaluate thermodynamic quantities as a function of temperature in the chosen range.[6] From the REMD simulation performed for this study, the properties of interest have been extracted at any temperature of the set and the aggregation of the amyloid-forming peptide GNNQQNY has been monitored in temperature space (275–400 K). This analysis gives interesting insights into the amyloid aggregation process.

The effective energy histograms shown in Fig. 2 are not symmetrically distributed around their mean value and their shape varies with temperature. The distributions, in fact, broaden toward higher energy values at low temperature (275–344 K) and toward lower energy values at high temperature (371–400 K). Moreover, by increasing temperature they progressively become lower and broader till the value of 371 K is reached. Mitsutake *et al.* have interpreted such a behavior as the evidence of a phase transition.[61] To characterize the transition, the radius of gyration R_g of the oligomeric system was considered and free-energy projections along R_g were plotted (see Fig. 5). Conformations of the system producing non-interacting peptides, namely, confor-

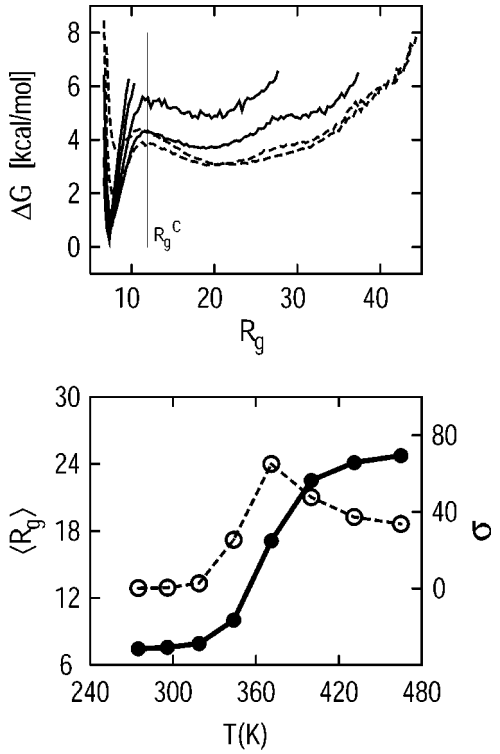


FIG. 5. (Top) Free-energy projections along the radius of gyration of the oligomeric system R_g computed from REMD trajectories. Solid lines correspond to temperature values below the condensation temperature (275–344 K); dashed lines correspond to temperature values above the transition temperature (371 and 400 K). The lowest radius of gyration for the uncondensed state is shown as a vertical line ($R_g^C = 11.9$ Å). (Bottom) Temperature dependence of the average radius of gyration $\langle R_g \rangle$ (filled circles) and its fluctuations σ (empty circles). The behavior of $\langle R_g \rangle$ and σ indicates the presence of a phase transition around 371 K between a condensed (low T) and an uncondensed phase (high T). Fluctuations of the radius of gyration σ are computed as $\langle R_g^2 \rangle - \langle R_g \rangle^2$. Data at 431 and 465 K were obtained from a preliminary REMD run carried out in a higher temperature range (6×1 μs, 319, 344, 371, 400, 431, and 465 K).

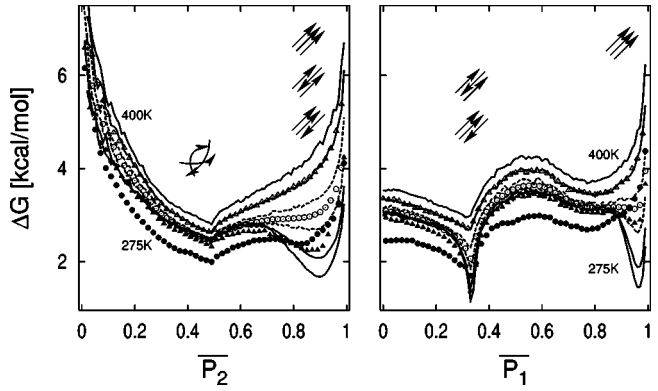


FIG. 6. Free-energy projections along the nematic ($\overline{P_2}$, left) and the polar ($\overline{P_1}$, right) order parameters. REMD data are shown in solid lines for all the temperature values except for 319 and 344 K which are in dashed lines. CTMD data are shown with symbols (filled circles, filled triangles, empty circles, and triangles for 275, 296, 330, and 371 K, respectively). Schematic representations of the aggregates (black arrows) are depicted to show that order parameters yield complementary information: $\overline{P_2}$ discriminates between ordered and disordered conformations while $\overline{P_1}$ discriminates between parallel and antiparallel/mixed ordered aggregates.

mations where all interpeptide atomic distances are larger than the long-range interactions cutoffs (7.5 Å in this case), were used to determine R_g^C, i.e., the lowest detected radius of gyration for isolated peptides (see Fig. 5). The existence of two macrostates in equilibrium has been revealed: the first, named *uncondensed state*, includes high energy conformations with one or more isolated peptides ($R_g > R_g^C$); the second, named *condensed state*, consists of low energy conformations with aggregated peptides ($R_g < R_g^C$). For entropic reasons, the *uncondensed state* is preferred at high temperature. By cooling down, the *condensed state* is increasingly stabilized, and around 371 K the fluctuations of R_g show a well-defined peak highlighting the presence of the *condensation* transition (see Fig. 5). The equilibrium between the *condensed* and the *uncondensed* macrostates is clearly concentration dependent. If the concentration of amyloid-forming units increases, the equilibrium is moved toward the condensed state and the aggregation process is favored.

The free-energy profiles along Q_p and Q_a at various temperatures help in understanding how the nucleation process evolves upon peptides condensation. At values of 400, 371, and 344 K both projections show steep uphill patterns with a single free-energy minimum at $Q_p \approx Q_a \approx 0$ (see Fig. 4). This means that upon condensation the peptides are still more likely to form disordered aggregates characterized by nonspecific interactions than amyloid-forming nuclei. In this range of temperatures, the enthalpic contribution due to in-register backbone or side-chain interactions does not dominate the entropic one and the growth of ordered nuclei is forbidden. However, when the temperature decreases the entropic contribution becomes less important and ordered in-register aggregates start forming. As shown in Fig. 4 in fact, below 330 K two and one additional free-energy minima appear in the projection along Q_p and Q_a, respectively. The observed minima correspond to in-register parallel ($Q_p > 0.7$) and in-register mixed or out-of-register ($0.3 \leqslant Q_p \leqslant 0.7$ and $0.4 \leqslant Q_a \leqslant 0.7$) arrangements and strongly suggest that the three-peptide system moves toward a higher degree of order when approaching the physiologically relevant conditions.

The simulation results indicate that in the early steps of amyloid aggregation a condensation stage toward disordered aggregates precedes the nucleation process and the disorder-order transition, in agreement with experimental evidence.[62]

C. Disorder-order transition

In the early steps of aggregation, amyloidogenic peptides assemble into highly ordered β-sheet structures.[21,30] During the assembly, the peptides tend to align adopting an extended β-strand conformation and a remarkable change in the local orientational order occurs. The aggregation of amyloid-forming peptides may then be interpreted as an order transition and orientational order parameters are suitable to monitor the time evolution of the process. Two orientational order parameters were employed and free-energy projections are shown in Fig. 6. Along $\overline{P_2}$, the free-energy profiles show a first broad minimum at $P_2 \approx 0.5$ for any temperature of the set and a second narrower one at $\overline{P_2} \approx 0.9$ for T values below

330 K. The first corresponds to a large free-energy basin where orientational order is absent, while the second corresponds to a smaller and well-defined basin with a high orientational degree of order. Although the order parameters should vanish when order is absent, Fig. 6 shows that this is not the case when the number of vectors is small. Since only three peptides were simulated, a ''background'' order was always detected and the free energy minimum describing the *disordered state* is placed at $\overline{P_2} \approx 0.5$, which is consistent with the value of $\sqrt{81/40\pi N}$ expected for a completely randomly oriented array of N molecules.[63] The order parameter $\overline{P_2}$ shows the existence of two macrostates in equilibrium: the disordered state with a high entropy content, which corresponds to the global minimum of the free energy surface at high temperature, and the *ordered state* which becomes the global free energy minimum at low temperature. Interestingly, the free-energy profiles along Q_p and Q_a do not lead to the same conclusion and the observed in-register arrangements correspond to local minima of the free-energy surface (see Fig. 4).

Along $\overline{P_1}$, two narrow and well-distinct minima corresponding to ordered macrostates at different polarity appear on the free-energy projections (Fig. 6). The first, displayed at $\overline{P_1} \approx 0.35$, describes a free-energy basin with a high-order and low-polarity content. Conversely, the second, displayed at $\overline{P_1} \approx 0.95$, corresponds to a basin with a high-order and high-polarity content. The order parameter $\overline{P_1}$ discriminates between parallel and antiparallel/mixed ordered conformations and provides complementary information since it allows to further characterize the ordered state.

Symbols in Fig. 6 show the free-energy projections along the order parameters from CTMD simulations. Once again, the comparison with REMD profiles indicates that isothermal MD (filled symbols) does not sample the ordered aggregates with their correct statistical weight close to the physiological temperature range.

The REMD free energy profiles along $\overline{P_1}$ show that at low temperature (275 and 296 K) both polar macrostates are highly populated. In the investigated temperature range, the system does not show an overall polar degree and frequent jumps between ordered states characterized by different polarity are observed. This suggests that below the order transition the equilibrium between polar macrostates might help amyloidogenic systems overcoming the entropy loss occurring during nucleation. In other words the growth of amyloid-forming nuclei might have an entropically favorable component due to the multiple ordered macrostates.

D. Sequence dependence of amyloidogenic propensity

Free-energy projections along the nematic order parameter $\overline{P_2}$ show how the equilibrium between the ordered and disordered state changes in temperature space (Fig. 6). Upon cooling, the statistical weight of the ordered state increases and the mean of the $\overline{P_2}$ distribution moves toward higher values. The value of $\langle \overline{P_2} \rangle$, where $\langle \cdots \rangle$ indicates the average over the canonical ensemble, is then related to the thermodynamic stability of the ordered state and could be used to measure the amyloidogenic propensity of the system. $\langle \overline{P_2} \rangle$ values computed at different temperatures from REMD trajectories of the amyloid-forming peptide GNNQQNY are shown in Fig. 7 with filled circles. At high temperature, the $\langle \overline{P_2} \rangle$ values are close to 0.5 because no orientational order is present, and the system does not show amyloidogenicity. By decreasing temperature, the amyloidogenic propensity grows and becomes increasingly larger until the order transition is completed. At physiologically relevant conditions, $\langle \overline{P_2} \rangle \approx 0.65$ and the system is highly amyloidogenic in agreement with experimental data.[21]

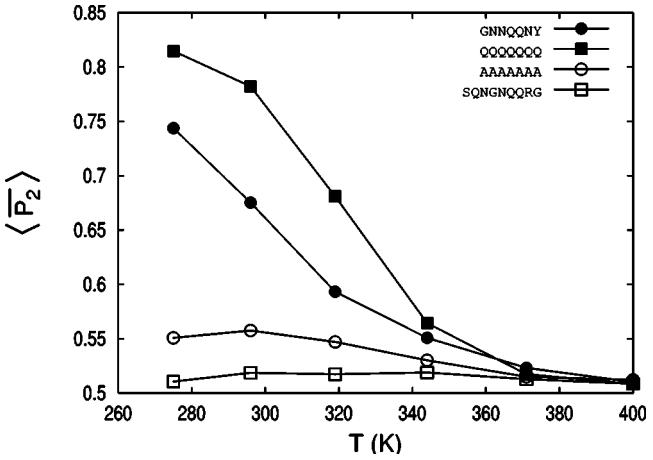


FIG. 7. Temperature dependence of the nematic order parameter $\langle \overline{P_2} \rangle$ averaged over the canonical ensembles sampled by REMD for four oligomeric peptide systems. $\langle \overline{P_2} \rangle$ estimates the amyloidogenic propensity of peptide systems and discriminates between amyloidogenic (GNNQQNY and QQQQQQQ) and nonamyloidogenic (SQNGNQQRG and AAAAAAA) sequences in agreement with experimental data (Refs. 21, 56, and 57).

Since the orientational order parameters do not depend on the peptide sequence and length, the reliability of the predictions could be further tested in sequence space. The REMD protocol was then applied to three additional oligomeric peptide systems (see Methods) and $\langle \overline{P_2} \rangle$ values were evaluated to measure and compare amyloidogenic propensities. The testing set comprises a nonapeptide from the yeast prion Sup35 (SQNGNQQRG) experimentally studied by Balbirnie *et al.*[21] and two heptapeptides (QQQQQQQ and AAAAAAA). Glutamine and alanine homopolymers flanked by basic residues to improve solubility have been investigated by Perutz *et al.*[56,57]

Experimentally, the nonapeptide SQNGNQQRG shows solubility *in vivo* and *in vitro* and no formation of amyloid fibrils.[21] In agreement with these findings, $\langle \overline{P_2} \rangle$ is smaller than 0.55 in the whole temperature range (Fig. 7, empty squares) and the system is considered as nonamyloidogenic. The number of aggregation events and the average lifetime of aggregation extracted from REMD trajectories are reported in Table I. Remarkably, these quantities show that nonamyloidogenic sequences, i.e, SQNGNQQRG and AAAAAAA, do transiently assemble in a β-sheet conformation but still remain soluble because their ordered aggregates do not correspond to well-defined free-energy minima.

Circular dichroism (CD) spectra, electron micrographs, and x-ray diffraction photographs showed that poly(L-

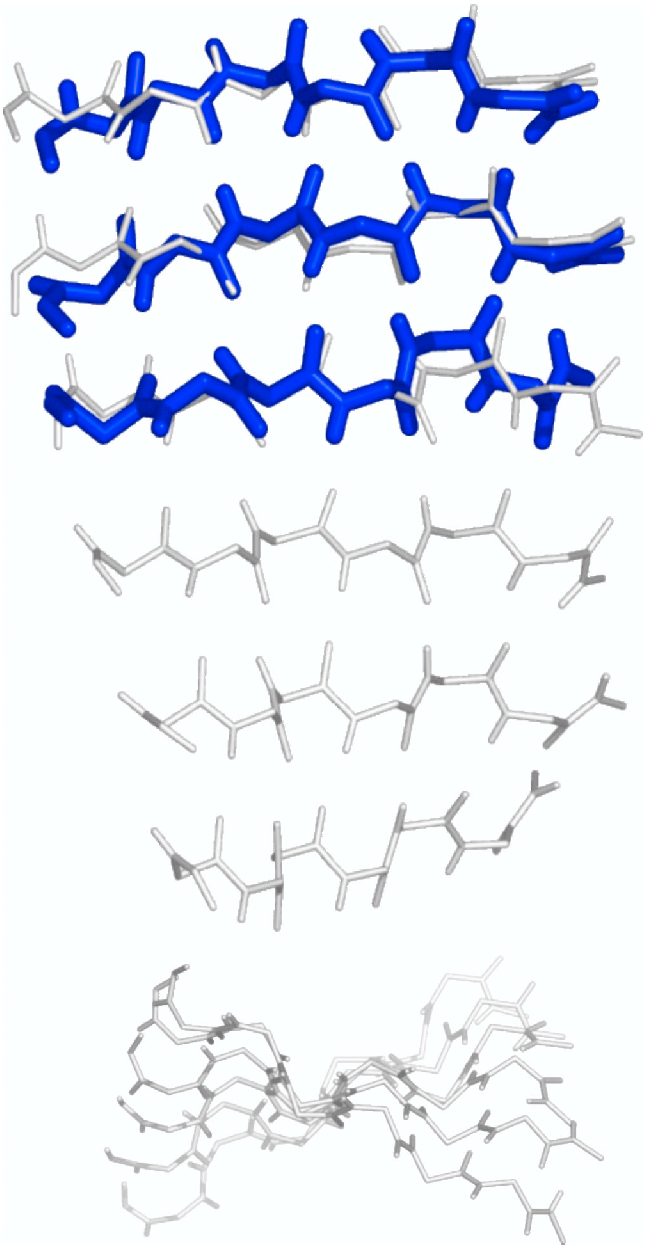


FIG. 8. (Top) Snapshots of ordered aggregates of three (thick sticks) and six (thin sticks) amyloidogenic SYVIIE peptides (Ref. 65) extracted from CTMD simulations at 330 K. The simulations were performed at a sample concentration of 5 mg/ml. The overall conformation and twist of the three-stranded and six-stranded parallel β-sheets are indistinguishable. (Bottom) The six-stranded β-sheet upon 90° rotation to better visualize the twist. [The pictures were drawn using the program PYMOL (Ref. 66)].

glutamine) peptides aggregate in solution at both *p*H 7.0 and 3.0 forming tightly linked β-sheet structures.[56] In particular, the x-ray diffraction picture exhibits a fiber diagram of the cross-β type distinctive of amyloid fibrils. On the other hand, poly(L-alanine) does not display amyloidogenicity and CD spectra showed α-helical structures at all *p*Hs.[57] Again, the $\langle \overline{P_2} \rangle$ patterns shown in Fig. 7 (filled squares and empty circles) are consistent with experimental findings and correctly indicate amyloidogenicity only for QQQQQQQ.

Interestingly, Fig. 7 allows also to compare between amyloidogenic sequences. In fact, according to the $\langle \overline{P_2} \rangle$ patterns the glutamine repeat is more amyloidogenic than GNNQQNY at physiologically relevant conditions. To our knowledge, no experimental data are available to verify this finding. Testing of this prediction is a challenge for experimentalists.

IV. CONCLUSIONS

The present study shows that atomistic REMD simulations with implicit solvent allow to sample the early steps of ordered aggregation of amyloidogenic peptides at physiologically relevant temperatures. The free-energy profiles projected along structural and orientational progress variables are essentially the same in REMD and CTMD. The discrepancies at temperature values below 330 K are due to the limitations in sampling in CTMD simulations which indicates that REMD is a more efficient approach in the physiological range.

The early steps of amyloidosis can be interpreted as a condensation followed by an order transition. Therefore, the REMD simulation results were analyzed with two order parameters originally introduced to study liquid crystals. Interestingly, the nematic order parameter averaged over a canonical ensemble is able to discriminate amyloidogenic from soluble peptides in agreement with experimental data.

Although the present study was performed with three peptides for reasons of computational efficiency, the description of the ordered aggregates is likely to be independent of the size of system, i.e., the number of simulated peptide replicas. Very recent MD simulations of the amyloidogenic SYVIIE peptide,[64] which has been experimentally investigated by de la Paz and Serrano,[65] have shown ordered aggregates of six peptides. Interestingly, the parallel β-sheet consisting of six peptides has the same overall conformation and twist as the three-peptide aggregate (Fig. 8).

ACKNOWLEDGMENTS

We thank Dr. U. Haberthür for running most of the CTMD simulations and R. Pellarin for introducing periodic boundary conditions in the SASA module in CHARMM (version 29). We are grateful to E. Guarnera and Dr. E. Paci for helpful discussions. We thank A. Widmer (Novartis Pharma, Basel) for providing the molecular modeling program WIT!P which was used for visual analysis of the trajectories. The simulations were performed on the Matterhorn Beowulf cluster at the Computing Center of the University of Zurich. We thank C. Bollinger and Dr. A. Godknecht for setting up the cluster and the Canton of Zurich for generous hardware support. This work was supported by the Swiss National Competence Center in Structural Biology (NCCR) and the Swiss National Science Foundation (Grant No. 31-64968.01 to A.C.).

[1] D. Frenkel and B. Smit, *Understanding Molecular Simulations* (Academic Press, San Diego, 2002).
[2] B. J. Berne and J. E. Straub, Curr. Opin. Struct. Biol. **7**, 181 (1997).
[3] A. Mitsutake, Y. Sugita, and Y. Okamoto, Biopolymers **60**, 96 (2001).
[4] N. Rathore, T. A. Knotts IV, and J. J. de Pablo, J. Chem. Phys. **118**, 4285 (2003).
[5] E. Marinari and G. Parisi, Europhys. Lett. **19**, 451 (1992).
[6] Y. Sugita and Y. Okamoto, Chem. Phys. Lett. **314**, 141 (1999).
[7] C. D. Michele and F. Sciortino, Phys. Rev. E **65**, 051202 (2002).

[8] K. Sanbonmatsu and A. Garcia, Proteins: Struct., Funct., Genet. **46**, 225 (2002).
[9] A. E. Garcia and K. Sanbonmatsu, Proc. Natl. Acad. Sci. U.S.A. **99**, 2782 (2002).
[10] A. E. Garcia and K. Sanbonmatsu, Proteins: Struct., Funct., Genet. **42**, 345 (2001).
[11] R. Zhou, B. Berne, and R. Germain, Proc. Natl. Acad. Sci. U.S.A. **98**, 14931 (2001).
[12] A. E. Garcia and J. N. Onuchic, Proc. Natl. Acad. Sci. U.S.A. **100**, 13898 (2003).
[13] J. W. Pitera and W. Swope, Proc. Natl. Acad. Sci. U.S.A. **100**, 7587 (2003).
[14] F. Rao and A. Caflisch, J. Chem. Phys. **119**, 4035 (2003).
[15] W. Im and C. L. Brooks, III, J. Mol. Biol. **337**, 513 (2004).
[16] C. M. Dobson, Trends Biochem. Sci. **24**, 329 (1999).
[17] M. F. Perutz, Trends Biochem. Sci. **24**, 58 (1999).
[18] C. Blake and L. Serpell, Structure (London) **4**, 989 (1996).
[19] S. B. Malinchik, H. Inouye, K. E. Szumowski, and D. A. Kirschner, Biophys. J. **74**, 537 (1998).
[20] M. Bucciantini, E. Giannoni, F. Chiti *et al.*, Nature (London) **416**, 507 (2002).
[21] M. Balbirnie, R. Grothe, and D. Eisenberg, Proc. Natl. Acad. Sci. U.S.A. **98**, 2375 (2001).
[22] R. A. Broglia, G. Tiana, S. Pasquali, H. E. Roman, and E. Vigezzi, Proc. Natl. Acad. Sci. U.S.A. **95**, 12930 (1998).
[23] D. Bratko and H. W. Blanch, J. Chem. Phys. **118**, 5185 (2003).
[24] G. Giugliarelli, C. Micheletti, J. R. Banavar, and A. Maritan, J. Chem. Phys. **113**, 5072 (2000).
[25] P. M. Harrison, H. S. Chan, S. B. Prusiner, and F. E. Cohen, J. Mol. Biol. **286**, 593 (1999).
[26] A. V. Smith and C. K. Hall, J. Mol. Biol. **312**, 187 (2001).
[27] B. Vekhter and R. S. Berry, J. Chem. Phys. **110**, 2195 (1999).
[28] M. Friedel and J. E. Shea, J. Chem. Phys. **120**, 5809 (2004).
[29] A. Fernandez and M. Boland, FEBS Lett. **529**, 298 (2002).
[30] J. Gsponer, U. Habertür, and A. Caflisch, Proc. Natl. Acad. Sci. U.S.A. **100**, 5154 (2003).
[31] E. Paci, J. Gsponer, X. Salvatella, and M. Vendruscolo, J. Mol. Biol. **340**, 555 (2004).
[32] F. Massi, J. W. Peng, J. P. Lee, and J. E. Straub, Biophys. J. **80**, 31 (2001).
[33] B. Ma and R. Nussinov, Protein Sci. **11**, 2335 (2002).
[34] B. Ma and R. Nussinov, Proc. Natl. Acad. Sci. U.S.A. **99**, 14126 (2002).
[35] D. Klimov and D. Thirumalai, Structure (London) **11**, 295 (2003).
[36] G. Tiana, F. Simona, R. A. Broglia, and G. Colombo, J. Chem. Phys. **120**, 8307 (2004).
[37] P. Ferrara, J. Apostolakis, and A. Caflisch, Proteins: Struct., Funct., Genet. **46**, 24 (2002).
[38] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, J. Comput. Chem. **4**, 187 (1983).
[39] E. Neria, S. Fischer, and M. Karplus, J. Chem. Phys. **105**, 1902 (1996).
[40] W. Hasel, T. F. Hendrickson, and W. C. Still, ''A Rapid Approximation to the Solvent Accessible Surface Arteas of Atoms,'' Tetrahedron Computer Methodology (Pergamon, New York, 1998), Vol. **1**, No. 2, pp. 103–116.
[41] T. Lazaridis and M. Karplus, Proteins: Struct., Funct., Genet. **35**, 133 (1999).
[42] A. Hiltpold, P. Ferrara, J. Gsponer, and A. Caflisch, J. Phys. Chem. B **104**, 10080 (2000).
[43] P. Ferrara and A. Caflisch, Proc. Natl. Acad. Sci. U.S.A. **97**, 10780 (2000).
[44] P. Ferrara and A. Caflisch, J. Mol. Biol. **306**, 837 (2001).
[45] J. Gsponer and A. Caflisch, Proc. Natl. Acad. Sci. U.S.A. **99**, 6719 (2002).
[46] J. Gsponer and A. Caflisch, J. Mol. Biol. **309**, 285 (2001).
[47] J. P. Hansen and I. R. McDonald, *Theory of Simple Liquids*, 2nd ed. (Academic Press, Oxford, 1990).
[48] J. P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen, J. Comput. Phys. **23**, 327 (1977).
[49] M. E. Rose, *Elementary Theory of Angular Momentum* (Wiley, New York, 1957).
[50] C. Zannoni, *The Molecular Physics of Liquid Crystals* (Academic, London, 1979), Chap. 3.
[51] S. Chandrasekhar, *Liquid Crystals* (Cambridge University Press, Cambridge, England, 1992).
[52] P. G. de Gennes and J. Prost, *The Physics of Liquid Crystals*, 2nd ed. (Oxford University Press, Oxford, 1993).
[53] C. Zannoni, J. Mater. Chem. **11**, 2637 (2001).
[54] R. Berardi, L. Muccioli, and C. Zannoni, ChemPhysChem **5**, 104 (2004).
[55] M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids* (Oxford Science, Oxford, UK, 1987).
[56] M. F. Perutz, T. Johnson, M. Suzuky, and J. T. Finch, Proc. Natl. Acad. Sci. U.S.A. **91**, 5355 (1994).
[57] M. F. Perutz, B. J. Pope, D. Owen, E. E. Wanker, and E. Scherzinger, Proc. Natl. Acad. Sci. U.S.A. **99**, 5596 (2002).
[58] M. Seeber, M. Cecchini, F. Rao, G. Settanni, and A. Caflisch (unpublished).
[59] W. Humphrey, A. Dalke, and K. Schulten, J. Mol. Graphics **14**, 33 (1996).
[60] M. K. Fenwick and F. A. Escobedo, J. Chem. Phys. **119**, 11998 (2003).
[61] A. Mitsutake, Y. Sugita, and Y. Okamoto, J. Chem. Phys. **118**, 6676 (2003).
[62] T. R. Serio, A. G. Cashikar, A. S. Kowal, G. J. Sawicki, J. J. Moslehi, L. Serpell, M. F. Arnsdorf, and S. L. Lindquist, Science **289**, 1317 (2000).
[63] T. P. Doerr, D. Herman, H. Mathur, and P. L. Taylor, Europhys. Lett. **59**, 398 (2002).
[64] Cecchini *et al.* (unpublished).
[65] M. Lopez de la Paz and L. Serrano, Proc. Natl. Acad. Sci. U.S.A. **101**, 87 (2004).
[66] W. DeLano, *The* PYMOL *Molecular Graphics System* (DeLano Scientific, San Carlos, CA, 2002).

Chapter 7

A Molecular Dynamics Approach to the Structural Characterization of Amyloid Aggregation

(Journal of Molecular Biology 357, pp 1306-1321, 2006)

doi:10.1016/j.jmb.2006.01.009 J. Mol. Biol. (2006) 357, 1306–1321

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

A Molecular Dynamics Approach to the Structural Characterization of Amyloid Aggregation

M. Cecchini[1], R. Curcio[1], M. Pappalardo[2], R. Melki[3] and A. Caflisch[1]*

[1]*Department of Biochemistry University of Zurich, Winterthurerstrasse 190 CH-8057 Zurich, Switzerland*

[2]*Dipartimento di Scienze Chimiche, Università di Catania Viale Andrea Doria 6, 95125 Catania, Italy*

[3]*Laboratoire d'Enzymologie et Biochimie Structurales, CNRS Avenue de la Terrasse, 91198 Gif-sur-Yvette, France*

**Corresponding author*

A novel computational approach to the structural analysis of ordered β-aggregation is presented and validated on three known amyloidogenic polypeptides. The strategy is based on the decomposition of the sequence into overlapping stretches and equilibrium implicit solvent molecular dynamics (MD) simulations of an oligomeric system for each stretch. The structural stability of the in-register parallel aggregates sampled in the implicit solvent runs is further evaluated using explicit water simulations for a subset of the stretches. The β-aggregation propensity along the sequence of the Alzheimer's amyloid-β peptide ($A\beta_{42}$) is found to be highly heterogeneous with a maximum in the segment $V_{12}HHQKLVFFAE_{22}$ and minima at S_8G_9, $G_{25}S_{26}$, $G_{29}A_{30}$, and $G_{38}V_{39}$, which are turn-like segments. The simulation results suggest that these sites may play a crucial role in determining the aggregation tendency and the fibrillar structure of $A\beta_{42}$. Similar findings are obtained for the human amylin, a 37-residue peptide that displays a maximal β-aggregation propensity at Q_{10}-$RLANFLVHSSNN_{22}$ and two turn-like sites at $G_{24}A_{25}$ and $G_{33}S_{34}$. In the third application, the MD approach is used to identify β-aggregation "hot-spots" within the N-terminal domain of the yeast prion Ure2p ($Ure2p_{1–94}$) and to design a double-point mutant ($Ure2p\text{-}N4748S_{1–94}$) with lower β-aggregation propensity. The change in the aggregation propensity of $Ure2p\text{-}N4748S_{1–94}$ is verified *in vitro* using the thioflavin T binding assay.

Keywords: amyloid; Alzhemer's disease; prion; protein aggregation; site-directed mutagenesis

Introduction

Protein folding and unfolding are the most sophisticated and specific ways of promoting and abolishing cellular activity. Failure to fold correctly, or to remain folded correctly, results in a plethora of diseases.[1–3] Some of these diseases originate from amyloidogenic polypeptides that have a high propensity to form ordered aggregates in the extracellular space,[4] e.g. Alzheimer's and Parkinson's diseases, type II diabetes, systemic amyloidosis, and transmissible spongiform encephalopathies.[5,6] Understanding the molecular determinants that cause soluble proteins to aggregate into insoluble fibrils and plaques is therefore an important challenge.

The soluble precursors of amyloid deposits do not share any sequence homology or common fold. However, amyloid aggregates have common structural features: (i) they show the same optical behavior (such as birefringence) on binding certain dye molecules such as Congo red; (ii) present very similar morphologies (long, unbranched and often twisted fibrillar structures a few nanometers in diameter); and (iii) display the characteristic cross-β X-ray diffraction pattern,[7,8] which indicates that the "core" structure is composed of β-sheets running perpendicular to the fibril axis.[9] Hence, regardless of the sequence, the key steps in the aggregation process may be common to all amyloidogenic polypeptides. The ability of a polypeptide chain to self-assemble is not restricted to disease-related proteins. A large number of non-pathogenic peptides and proteins have been shown to form amyloid fibrils under particular solvent, pH and temperature conditions.[10–12] Taken together, these

Abbreviations used: MD, molecular dynamics; APP, amyloid precursor protein; REMD, replica exchange MD; ThT, thioflavin T; EPR, electron paramagnetic resonance; SS, secondary structure; hIAPP, human islet amyloid polypeptide; Aβ42, Alzheimer's amyloid-β peptide.

E-mail address of the corresponding author: caflisch@bioc.unizh.ch

observations indicate that amyloid propensity is a general property of the polypeptide backbone,[13] thus suggesting that under certain conditions any protein above a critical concentration will eventually assemble into ordered aggregates. On the other hand, several studies have shown that the aggregation propensity depends dramatically on amino acid composition and that side-chains influence the kinetics and stability of amyloid fibrils enormously.[14–17] To shed light into the molecular mechanism of amyloid formation and the nature of the energetic contributions that stabilize these structures for an extremely diverse class of polypeptides, atomic-resolution three-dimensional structures are required. As non-crystalline solid material, amyloid aggregates are strongly incompatible with high-resolution techniques for protein structure determination, i.e. X-ray crystallography and liquid state NMR, and, with the exception of the amyloid-like spine formed by a seven-residue peptide,[18] no structure of an amyloid fibril has yet been determined at an atomic level of detail. To obtain structural information, more sophisticated approaches such as solid state NMR,[19–24] site-directed spin labeling,[25–27] cryo-electron microscopy[28,29] and proline-scanning mutagenesis[30] have been used.

Given the difficulty of obtaining high-resolution structures, alternative theoretical and computational approaches have been followed to rationalize the physico-chemical principles of amyloidogenesis and understand the role of the sequence. Very efficient theoretical models to predict protein aggregation propensities from primary structures have been proposed.[31–33] At minimal computational cost, some of these models[32,33] determine putative aggregation-prone regions ("hot-spots") within a protein sequence. Despite remarkable correlation with experimental data, these methods do not provide detailed structural information. Experimental approaches on simplified amyloid systems have also been reported.[17] Remarkably, the full positional scanning mutagenesis of the amyloidogenic peptide STVIIE highlighted both sequence and position dependence of amyloid propensity, even for such a small peptide system. However, the lack of structural detail for the fibrils formed by these peptides has precluded a rational explanation for the origin of the observed mutational effects. Hence, to shed some light into the "hidden" link between protein sequence and amyloid propensity, computational studies providing structural information are required.[34,35]

Here, a novel approach to structurally characterize the propensity towards ordered aggregation of amyloid polypeptides is presented. The procedure is based on the decomposition of a polypeptide chain into overlapping segments and equilibrium molecular dynamics (MD) simulations of a small number of copies of each segment. An efficient implicit solvent model (based on the solvent-accessible surface area[36]) is used to obtain a statistically significant sampling for the trimeric and hexameric systems of each peptide segment. It is important to note in this context that the in-register parallel packing of a seven-residue peptide from the yeast prion protein Sup35 determined by this implicit solvent model[16] is in remarkable agreement with the X-ray microcrystal structure of the cross-β spine (see Materials and Methods).[18] The computational strategy has been designed to predict the position dependence of β-aggregation propensity along the sequence, i.e. the amyloidogenicity profile. From the shape of the profile, amyloidogenic stretches can be discriminated from regions with scarce β-aggregation propensity. Moreover, the atomic detail provided by the MD simulations allows us to interpret the shape of the profile on a structural basis. The method has been tested on three amyloid sequences: the amyloid-β peptide ($A\beta_{42}$), the human amylin and the N-terminal domain of the yeast prion protein Ure2 ($Ure2p_{1–94}$). $A\beta_{42}$ is a product of the proteolytic cleavage of the 695-residue amyloid precursor protein (APP) accomplished by the β and γ-secretases.[37] Amyloid fibrils and plaques formed by full-length Aβ are associated with Alzheimer's disease, which is the most common neurodegenerative disease and accounts for the majority of the dementia diagnosed after the age of 60.[38]

Human amylin, also known as islet amyloid polypeptide (hIAPP), is the major component of pancreatic amyloid deposits found in ~90% of patients with non-insulin-dependent (type 2) diabetes mellitus[39] of which there are about 150 million worldwide.[40] hIAPP is a peptide hormone of 37 amino acid residues produced by cleavage from a pro-amylin precursor protein. It has been shown by X-ray and electron diffraction that hIAPP fibrils are well-ordered cross-β structures.[41] However, a detailed understanding of the fibrillar structure and aggregation properties of full-length hIAPP ($hIAPP_{1–37}$) has yet to be achieved.

Ure2p is a prion from the yeast *Saccharomyces cerevisiae*[42] that acts as a negative regulator of nitrogen metabolism.[43] In its prion state, Ure2p is at the origin of the [URE3] phenotype.[44] *In vitro*, Ure2p aggregates into long, straight filaments that bind Congo red, show green-yellow birefringence and have an increased resistance to proteolysis.[45,46] The prion domain (residues 1–90)[46] is involved in filament formation[47] and contains an unusually high number of Asn, Gln, and Ser residues, i.e. 35%, 12%, and 10% of its 90 residues, respectively. The prion domain of Ure2p is protease-sensitive and poorly structured.[46] However, synthetic $Ure2p_{1–65}$ was shown to readily form fibrils with more than 60% β-sheet.[47]

The agreement between the amyloidogenicity profile obtained by MD simulations and experimental data on $A\beta_{42}$ and $hIAPP_{1–37}$ indicates that the computational approach is descriptive and can be applied to predict the aggregation properties of a polypeptide sequence. The β-aggregation profile highlights critical segments for β-sheet formation

and can be used to guide site-directed mutations that modulate the aggregation tendency. The predictive power of the computational approach is validated experimentally by a double-point mutant of $Ure2p_{1–94}$.

Results

Polypeptide decomposition into segments

Here, the aggregation properties of three amyloid sequences are investigated by the MD computational approach: the human $A\beta_{42}$, $hIAPP_{1–37}$, and the central part of the N-terminal domain of the yeast prion protein Ure2 ($Ure2p_{20–70}$). Due to the large size of the polypeptide chains (42, 37 and 51 residues, respectively) their oligomeric systems cannot be effectively studied by all-atom MD simulations. Therefore, the polypeptide sequence was decomposed into overlapping stretches. By systematically applying a two-residue shift along the sequence, 18 seven-residue and 16 11-residue peptide segments span the 1–41 region of $A\beta_{42}$ (Tables 1 and S2), 16 seven-residue peptide segments span $hIAPP_{1–37}$ (Table S3), and 23 seven-residue peptide segments cover the 20–70 region of Ure2p (Table S4). Each peptide segment was both N-acetylated and C-amidated to reproduce the original context in the full-length sequence. The considerable overlap between neighboring segments allows us to extrapolate the simulation results from the stretches to the polypeptides.

Human amyloid-β peptide ($A\beta_{42}$)

β-Aggregation propensity

The β-aggregation profile of $A\beta_{42}$ was determined by first performing implicit solvent MD simulations of a trimeric system for each of the seven-residue peptide segments (see Table 1). For each segment, the β-aggregation propensity was obtained by averaging the time series of the nematic order parameter $\bar{P}_2$ along the trajectory (see Materials and Methods). At both temperatures of 310 K and 330 K the average value of $\bar{P}_2$ reaches convergence, on a time-scale faster than 1 μs, as indicated by the small standard deviations from three independent runs at 310 K (Figure 1, top). At 330 K, β-aggregation propensity values ranged from 0.51 to 0.89. According to our previous analysis of the amyloid-forming peptides GNNQQNY and QQQQQQQ and nonamyloidogenic peptides SQNGNQQRG and AAAAAAA,[48] these $\bar{P}_2$ values indicate the presence of both aggregation-prone and non aggregation-prone stretches along the $A\beta_{42}$ sequence. Interestingly, the most aggregation-prone segments are not distributed uniformly along the $A\beta_{42}$ sequence but tend to cluster in the region 12–22. The heterogeneity in the aggregation properties of the $A\beta_{42}$ segments is reflected in the free-energy projections along $\bar{P}_2$ (Figure 1, bottom left). Two main scenarios emerge: the first, described by a free-energy profile with a broad minimum at $\bar{P}_2 \sim 0.5$ (broken line),

Table 1. $A\beta_{42}$: seven-residue stretches simulations

Segment	Peptide sequence	Central Residue	Three peptides (μs) 310 K	Three peptides (μs) 330 K	Six peptides (μs) 330 K
$A\beta_{1-7}$	**DAEFRHD**SGYEVHHQKLVFFAEDVGSNKGAIIGLMVGGVVIA	4	3 × 1.0	1 × 1.4	1 × 1.5
$A\beta_{3-9}$	DA**EFRHDSG**YEVHHQKLVFFAEDVGSNKGAIIGLMVGGVVIA	6	3 × 1.0	1 × 1.3	1 × 1.6
$A\beta_{5-11}$	DAEF**RHDSGYE**VHHQKLVFFAEDVGSNKGAIIGLMVGGVVIA	8	3 × 1.0	1 × 1.1	1 × 1.6
$A\beta_{7-13}$	DAEFRH**DSGYEVH**HQKLVFFAEDVGSNKGAIIGLMVGGVVIA	10	3 × 1.0	1 × 1.6	1 × 1.8
$A\beta_{9-15}$	DAEFRHDS**GYEVHHQ**KLVFFAEDVGSNKGAIIGLMVGGVVIA	12	3 × 1.0	1 × 1.4	1 × 1.7
$A\beta_{11-17}$	DAEFRHDSGY**EVHHQKL**VFFAEDVGSNKGAIIGLMVGGVVIA	14	3 × 1.0	1 × 1.3	1 × 1.7
$A\beta_{13-19}$	DAEFRHDSGYEV**HHQKLVF**FAEDVGSNKGAIIGLMVGGVVIA	16	3 × 1.0	1 × 1.9	1 × 1.6
$A\beta_{15-21}$	DAEFRHDSGYEVHH**QKLVFFA**EDVGSNKGAIIGLMVGGVVIA	18	3 × 1.0	1 × 1.7	1 × 1.8
$A\beta_{17-23}$	DAEFRHDSGYEVHHQK**LVFFAED**VGSNKGAIIGLMVGGVVIA	20	3 × 1.0	1 × 1.3	1 × 2.0
$A\beta_{19-25}$	DAEFRHDSGYEVHHQKLV**FFAEDVG**SNKGAIIGLMVGGVVIA	22	3 × 1.0	1 × 1.3	1 × 2.0
$A\beta_{21-27}$	DAEFRHDSGYEVHHQKLVFF**AEDVGSN**KGAIIGLMVGGVVIA	24	3 × 1.0	1 × 1.5	1 × 2.3
$A\beta_{23-29}$	DAEFRHDSGYEVHHQKLVFFAE**DVGSNKG**AIIGLMVGGVVIA	26	3 × 1.0	1 × 1.3	1 × 2.3
$A\beta_{25-31}$	DAEFRHDSGYEVHHQKLVFFAEDV**GSNKGAI**IGLMVGGVVIA	28	3 × 1.0	1 × 1.0	1 × 2.5
$A\beta_{27-33}$	DAEFRHDSGYEVHHQKLVFFAEDVGS**NKGAIIG**LMVGGVVIA	30	3 × 1.0	1 × 1.0	1 × 2.5
$A\beta_{29-35}$	DAEFRHDSGYEVHHQKLVFFAEDVGSNK**GAIIGLM**VGGVVIA	32	3 × 1.0	1 × 1.1	1 × 2.8
$A\beta_{31-37}$	DAEFRHDSGYEVHHQKLVFFAEDVGSNKGA**IIGLMVG**GVVIA	34	3 × 1.0	1 × 1.1	1 × 2.8
$A\beta_{33-39}$	DAEFRHDSGYEVHHQKLVFFAEDVGSNKGAII**GLMVGGV**VIA	36	3 × 1.0	1 × 1.2	1 × 3.2
$A\beta_{35-41}$	DAEFRHDSGYEVHHQKLVFFAEDVGSNKGAIIGL**MVGGVVI**A	38	3 × 1.0	1 × 1.2	1 × 3.2

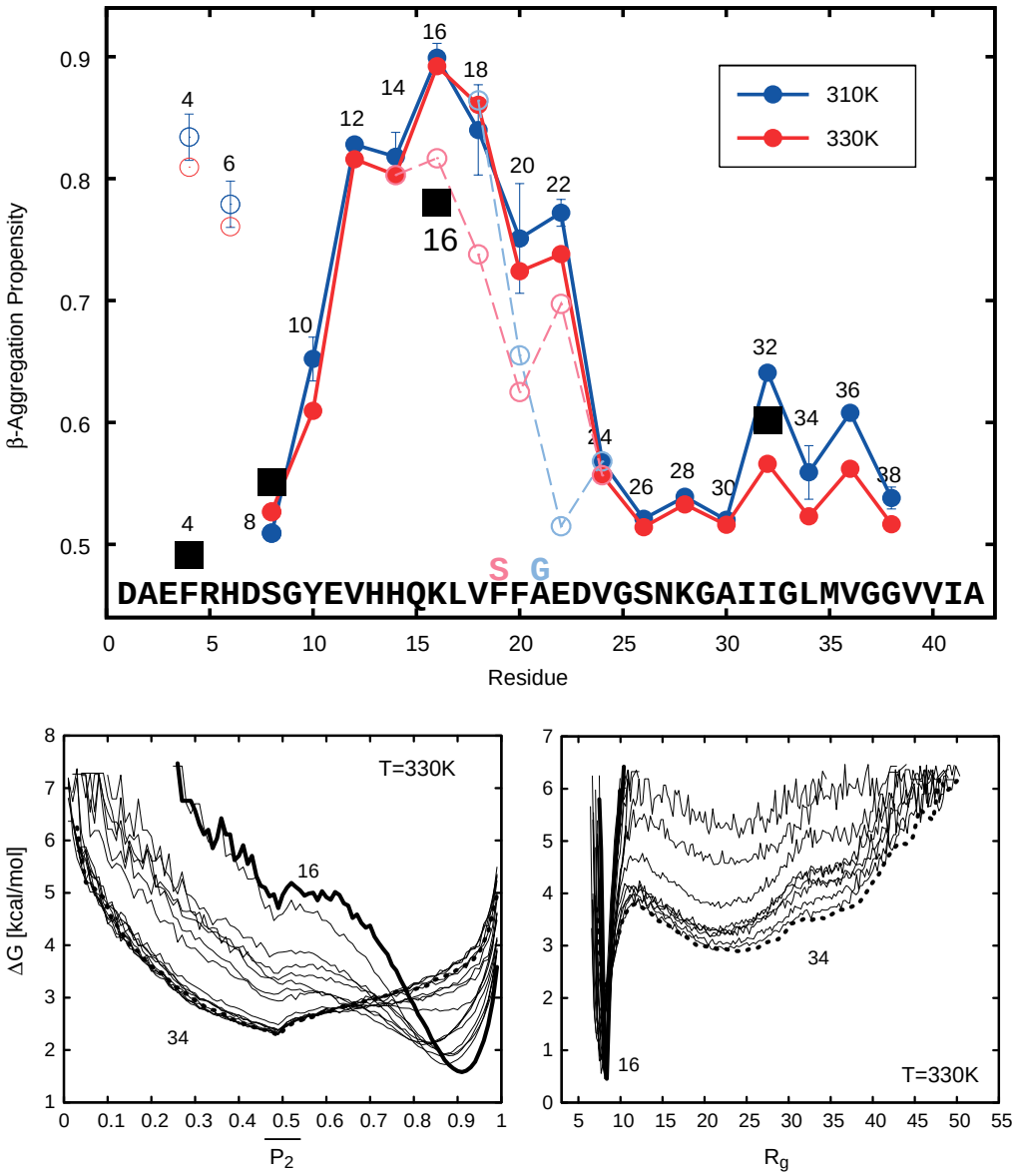


Figure 1. Aβ$_{42}$. Results of constant temperature MD simulations of trimeric seven-residue peptide systems. Top: β-aggregation propensity averaged along the implicit solvent simulations at 310 K (blue), 330 K (red), and the explicit solvent runs (black squares). The data points represent the values of the nematic order parameter $\bar{P}_2$ averaged over the canonical ensemble. The continuous and broken lines are drawn to help the eye for the wild-type and single-point mutants, respectively. The segment identification number corresponds to the position of the central residue in the Aβ$_{42}$ full-length sequence (see Table 1). Error bars on the data points at 310 K represent standard deviations of average β-propensities computed over three independent runs. For some data points at 310 K, e.g. 32 and 36, the error bar is smaller than the symbol. Pink and cyan open circles show the effect of the single-point mutations, F19S[51] and A21G,[52] respectively. Bottom: free-energy projections along $\bar{P}_2$ (left) and the radius of gyration (right) of the oligomeric system at 330 K. Thick and broken lines for segments 16 and 34, respectively, show the emergence of two distinct scenarios in both aggregation (left) and condensation (right) properties. Thin continuous lines represent the 16 remaining segments.

indicates scarce propensity for β-aggregation; the second, described by a steep downhill profile toward a minimum with high orientational order (thick line), highlights amyloid-like sequences.

To obtain insights into the relatively weak β-aggregation propensity detected at the C terminus, the radius of gyration of the oligomeric system R_g was monitored along the implicit solvent trajectories. Again, the free-energy projections along R_g (Figure 1, bottom right) show two different scenarios. The first scenario, described by a free-energy profile with a unique and narrow minimum at $R_g < R_g^C$ (see Materials and Methods), indicates that the simulated peptides are very likely

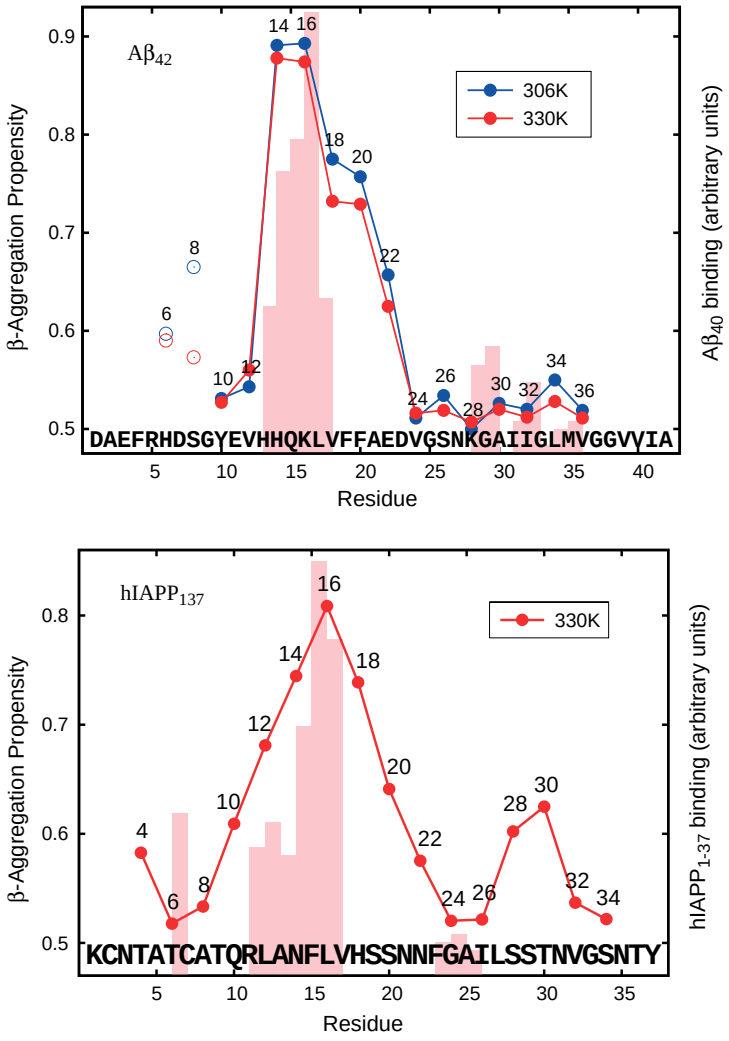


Figure 2. Comparison between β-aggregation propensities from MD simulations and experimental data. See the legend to Figure 1 for the meaning of data points and connecting lines. Top: 11-residue peptide segments of the amyloid-β peptide. The pink bars quantify the binding of the full-length $A\beta_{40}$ to each of 31 overlapping decapeptides (corresponding to residues 1–10 up to 31–40) as measured by radioligand experiments.[49] Bottom: seven-residue peptide segments of the human amylin. The pink bars quantify the binding of the full-length $hIAPP_{1-37}$ to each of 28 over lapping decap peptides (corresponding to residues 1–10 up to 28–37) as measured by immunoblotting experiments.[55]

to aggregate. The second scenario, which shows a prominent broad minimum at $R_g < R_g^C$ on the free-energy profile (broken line), reports that conformations characterized by isolated peptides are favored and the occurrence of condensed states of the system (either ordered or disordered) is rather low. Remarkably, all the stretches compatible with the second scenario belong to the C-terminal part of the amyloid-β peptide (from $A\beta_{21-27}$ to $A\beta_{35-41}$). Since at elevated temperature entropic effects favor the uncondensed state, three additional 1 μs implicit solvent simulations were performed at 310 K for each segment. Interestingly, an aggregation-prone region in the C-terminal part of $A\beta_{42}$ (residues 32–36) emerges at 310 K (Figure 1, blue circles). These simulation results suggest that the condensation propensity, though not sufficient to describe amyloidogenicity, is a necessary condition for the formation of amyloid nuclei. It is not sufficient because amyloidogenic sequences must also have β-sheet propensity, which promotes the assembly into highly ordered structures (see also Supplementary Data).

A comparison of the β-aggregation propensity calculated from the implicit and explicit solvent simulations shows a good agreement except for the N-terminal segment $D_1AEFRHD_7$ (Figure 1, top). The much larger $\bar{P}_2$ value in the implicit solvent runs is a consequence of the approximations inherent to the treatment of charged groups,[36] which are neutralized to prevent vacuum-like artifacts like the excessive formation of salt-bridges. In the implicit solvent simulations of $D_1AEFRHD_7$, the lack of strong Coulombic repulsion between side-chains with the same charge does not prevent formation of in-register parallel β-sheets. The first two N-terminal segments $D_1AEFRHD_7$ and E_3-$FRHDSG_9$ contain four and three charged side-chains, respectively, whereas the remaining seven-residue segments have between zero and two formal charges. For this reason, the implicit solvent β-aggregation profiles are more reliable for the 8–42

region of the $A\beta_{42}$ peptide, where they also show good agreement with the explicit solvent simulation results.

Implicit solvent replica exchange MD (REMD)[48] simulations of 11-residue segments were also used to derive the $A\beta_{42}$ amyloidogenicity profile (Figure 2, top). At 306 K and 330 K, the profiles look similar with higher propensity for the lower temperature. For all the aggregation-prone segments (8, 14, 16, 18, 20 and 22), the analysis of the trajectories by a polar order parameter ($\bar{P}_1$; see Cecchini *et al.*[48]) revealed a statistically relevant predominance of in-register parallel β-sheets and, with the exception of stretch 22, negligible anti-parallel arrangements (data not shown). The high-propensity region encompasses residues 14–22. Interestingly, the same region was identified by means of radioligand experiments[49] as the most prone to bind full-length $A\beta_{42}$ (pink bars in Figure 2). The radioligand experiments were carried out with overlapping ten-residue stretches, which is very close to the simulation systems. A second region located at the C terminus, which is missing from the 11-residue segments aggregation profile, was detected by the experiments. However, the binding of $A\beta_{42}$ to the C-terminal decapeptides was considerably less prominent and probably mediated by hydrophobic rather than specific interactions.

An in-depth comparison of the amyloidogenicity profiles obtained from seven and 11-residue peptide simulations (Figures 1 and 2, respectively) provides additional information. The profiles are qualitatively similar and display a prominent high-propensity region located in the central part of the sequence. However, the effect of the increased peptide length is not negligible; a considerably lower β-aggregation propensity is detected at the N terminus (residues 1–6). When the peptide length is increased, the resulting stretches are more likely to include both aggregation-prone and non-aggregation-prone segments. Therefore, the β-aggregation propensity decreases in regions of the sequence with mixed properties. To verify that the length of the segments identified on the $A\beta_{42}$ sequence had no impact on the results, the β-aggregation propensity profile was recalculated by considering seven-residue substretches along the 11-residue simulation trajectories (e.g. D_1AEFRHD$_7$, E_3-FRHDSG$_9$ and R_5HDSGYE$_{11}$ from D_1-AEFRHDSGYE$_{11}$). Both β-aggregation propensity and secondary structure profiles are in good agreement with those obtained from seven-residue peptide simulations (Figure S4 in Supplementary Data). Thus, the simulation results are robust with respect to the choice of the segment length.

Secondary structure analysis

To interpret the amyloidogenic trend in the central zone (stretches 14, 16, 18, 20 and 22), a secondary structure analysis of the conformations saved along the trajectories was performed. The analysis showed that β-aggregation propensity correlates with β-strand content, anticorrelates with α-helical content, and seems very sensitive to β-turn or bend propensity (Figure 3 and Figure S5 in Supplementary Data). The β-aggregation profile in the segment 14–22 is influenced by the location of a turn-like segment ($G_{25}S_{26}$) and the α-helical/β-strand equilibrium. The latter is consistent with NMR studies that highlighted a trend to helical structures for the segment 16–24 in aqueous solution.[50] The identification of four turn-like

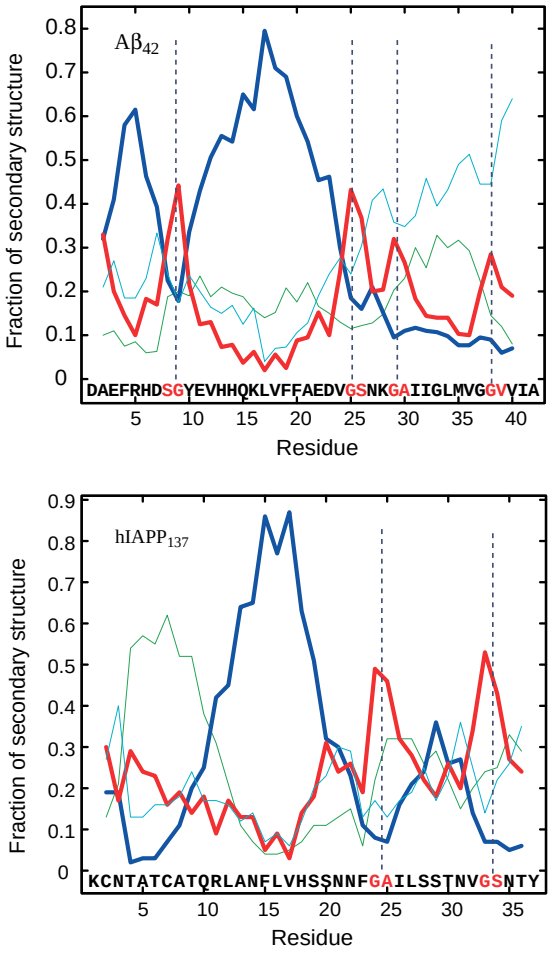


Figure 3. Secondary structure profiles. Single-residue secondary structure propensities have been extracted from REMD trajectory segments at 306 K of trimeric 11-residue systems for $A\beta_{42}$ (top) and from constant temperature MD trajectories at 330 K of trimeric seven-residue systems for $hIAPP_{1-37}$ (bottom). To obtain the propensity value of residue *i*, averages were taken over all stretches containing residue *i*. Green, blue, red, and cyan lines correspond to α-helical, β-strand, turn or bend, and random coil content, respectively. The simulation results indicate the presence of turn-like segments (red letters), which play a key role in determining amyloid aggregation properties. The vertical broken lines mark the borders of the regions identified by the specific location of the turns along the sequence.

peaks (S_8G_9, $G_{25}S_{26}$, $G_{29}A_{30}$, and $G_{38}V_{39}$) helps to interpret the shape of both seven and 11-residue profiles as well as the differences between the two. When $A\beta_{42}$ is dissected into 11-residue segments, the stretches at the N terminus always include the first potential turn (S_8G_9) in the middle of the sequence and therefore the amyloidogenic content is low. In contrast, the heptamer $A\beta_{3-9}$ has the S_8G_9 turn at the C terminus and shows a high β-aggregation propensity. Similar considerations explain the low propensity of the 11-residue profile in the C-terminal region of $A\beta_{42}$. With longer segments, either the third ($G_{29}A_{30}$) or the fourth ($G_{38}V_{39}$) turn-like site is always included in the stretches and the β-aggregation propensity is suppressed. On the other hand, the seven-residue stretches between $G_{29}A_{30}$ and $G_{38}V_{39}$ are responsible for the peak at residues 31–37 (Figure 1 and Figure S2 in Supplementary Data). Taken together, the simulation results show that the β-aggregation profile of $A\beta_{42}$ is strongly modulated by the position of four turn-like sites along the sequence.

Single-point mutants of Aβ 42

The aggregation properties of four familial disease-related variants of $A\beta_{42}$, i.e., the Arctic (E22G), Dutch (E22Q), Italian (E22K) and Flemish (A21G) mutants, and one non-pathological variant obtained by random mutation[51] (F19S) were investigated. *In vitro* studies[51–53] have shown that the E22G, E22Q and E22K mutations accelerate fibril formation while A21G and F19S decrease the fibrillogenesis rate with respect to wild-type $A\beta_{42}$. Starting from the β-aggregation profile of the wild-type sequence (Figure 1), mutational effects can be predicted at moderate computational cost, i.e. by repeating only the implicit solvent simulations of the seven-residue stretches affected by the mutation (Table S1 in Supplementary Data). Seven-residue segments are preferred to 11-residue segments because of the lower computational cost, which allows us to investigate a larger number of mutants.

As shown in Figure 1, the lower aggregation propensity of the F19S and A21G variants is reproduced correctly. Interestingly, the disease-related mutant E22G has a profile (data not shown) similar to that of wild-type $A\beta_{42}$, indicating that the simulation-based approach is able to distinguish the subtle difference between A21G and E22G. On the other hand, the fact that three disease-related variants E22G, E22Q and E22K have profiles similar to wild-type (data not shown) suggests that the approach is less sensitive in the very-high propensity region, which could be a consequence of the reduced dimensionality of the simulation system, i.e. number of peptides smaller than in the nucleus and/or short segment length. In this context it is important to note that previously published explicit water simulations of the monomeric $A\beta_{10-35}$ peptide and its E22Q mutant do not support the hypothesis that the Dutch E22Q variant leads to a higher β-structure propensity,[54] in agreement with the present implicit solvent results.

Human amylin ($hIAPP_{1-37}$)

The β-aggregation profile of $hIAPP_{1-37}$ was determined using the same approach as for $A\beta_{42}$. Sixteen implicit solvent MD simulations of three seven-residue peptide segments were performed (Table S3 in Supplementary Data). At 330 K, β-aggregation propensity values ranged from 0.52 to 0.81 (Figure 2, bottom). The resulting profile highlights two well-defined hot-spots along the $hIAPP_{1-37}$ sequence, with the first (residues 10–22) more prominent than the second (residues 28–30).

A systematic mapping of the $hIAPP_{1-37}$ sequence for the identification of domains that can potentially mediate molecular recognition and lead to amyloid fibril formation has been performed by Gazit and co-workers.[55] Their *in vitro* immunoblotting experiments showed that the region most prone to bind the full-length $hIAPP_{1-37}$, i.e. the recognition domain, is located at the center of the sequence (residues 7–21). The simulation results are in good agreement with *in vitro* findings (Figure 2, bottom). In particular, the NFVLH pentapeptide suggested as the "core" of the recognition motif[55] is included in the seven-residue stretch $hIAPP_{13-19}$ (central residue 16) that showed the highest β-aggregation propensity *in silico*. The largest discrepancy is located at the region 24–26, which was not identified by the simulation-based approach. However, the immunoblotting signal is very weak in this region and likely to originate from hydrophobic rather than specific interactions.

The structural analysis performed on $A\beta_{42}$ was repeated on the $hIAPP_{1-37}$ sequence. Average single-residue secondary structure propensities were extracted from simulation trajectories and used to draw the profiles shown in Figure 3, which highlights two short turn-like segments ($G_{24}A_{25}$ and $G_{33}S_{34}$) corresponding to regions of reduced β-aggregation propensity. Again, the simulation results suggest that turn-like sites strongly modulate the aggregation propensity of amyloid polypeptides. It is worth noting, though that a third β-aggregation propensity minimum observed around Thr6 is due to a strong α-helical propensity (green line in Figure 3 bottom) and not to a turn- or bend-site. As mentioned above, the α-helical/β-strand equilibrium can modulate the β-aggregation propensity of a polypeptide chain.

N-terminal domain of the prion protein Ure2 ($Ure2p_{1-94}$)

As for $A\beta_{42}$ and $hIAPP_{1-37}$, the β-aggregation profile of $Ure2p_{20-70}$ was determined by performing implicit solvent MD simulations of a trimeric system for each of the seven-residue peptide segments (Table S4 in Supplementary Data). At 330 K, β-aggregation propensity values ranged from 0.51 to 0.76 (Figure 4, top). Three aggrega-

tion-prone regions intercalated with short non aggregation-prone segments are highlighted: residues 33–39, 45–49, and 55–61. It is worth noting that one of the two central regions with low β-aggregation propensity corresponds to three consecutive serine residues located at positions 51–53. Apparently, these serine residues reduce the local aggregation tendency by splitting a poly (N) stretch in two segments. To further investigate the role of serine residues, six variants of the stretch Ure2p$_{44–50}$ (NNNNNNN) were modeled by considering all possible single and double-point N-to-S mutants at positions 47, 48 and 49 (Table S5 in Supplementary Data). Six additional simulations were run at 330 K and β-aggregation propensities computed. As shown in Figure 4 (top, blue triangles), a strong position dependence on mutation is observed in agreement with recent experimental findings.[17] Furthermore, the central positions (residues 47 and 48) are more sensitive than the lateral ones (residue 49). The N-to-S mutations reduce the aggregation propensity of Ure2p$_{44–50}$ with the lowest tendency for the double mutant Ure2p-N4748S$_{44–50}$. To study the effect of this double-point mutation on Ure2p$_{1–94}$, aggregation simulations of all stretches affected by the mutations, i.e. Ure2p$_{42–48}$, Ure2p$_{44–50}$, Ure2p$_{46–52}$ and Ure2p$_{48–54}$, were carried out. The N4748S double-point mutation is responsible for the disappearance of the hot-spot at residues 45–49 (empty circles in Figure 4 top) and therefore is predicted to strongly affect the assembly behavior of the whole prion domain. To validate the *in silico* prediction, the assembly kinetics of the N-terminal domain of wild-type Ure2p (Ure2p$_{1–94}$) and double-point mutant (Ure2p-N4748S$_{1–94}$) were compared using the thioflavin T (ThT) binding assay. For the latter, a pronounced increase in the lag phase of the assembly reaction and a lower level of ThT fluorescence at steady state were observed (Figure 4, bottom). From this observation, we conclude that the substitution of asparagine by serine residues at position 47 and 48 hinders the

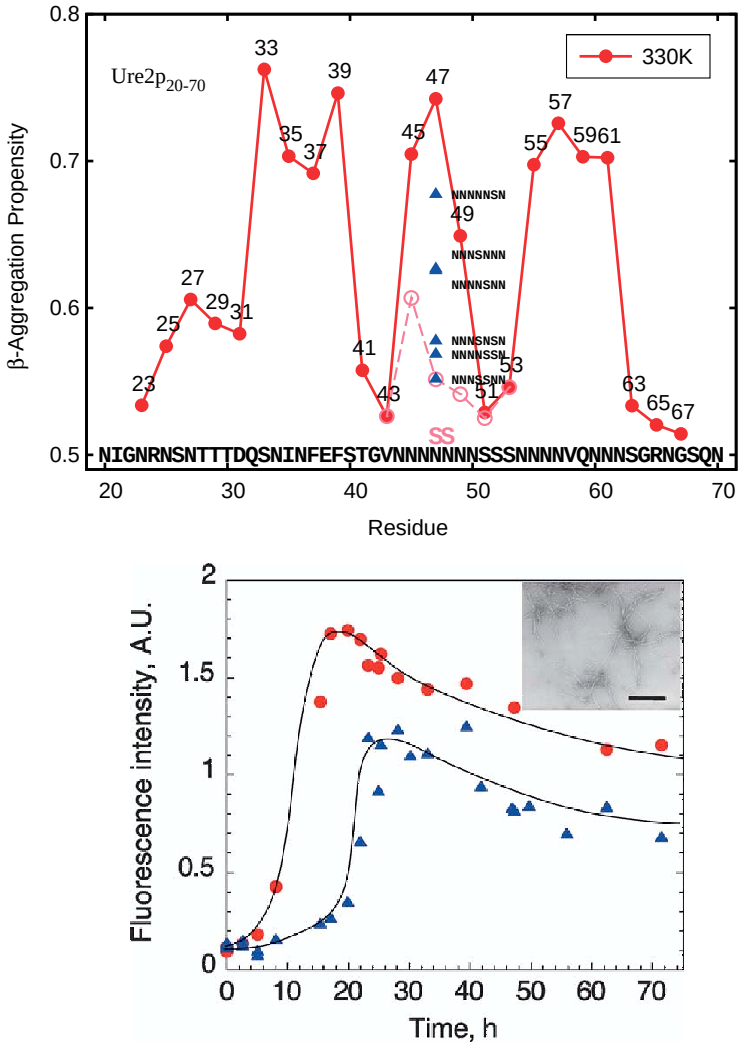


Figure 4. N-terminal domain of the prion protein Ure2: MD simulations and *in vitro* validation. Top: values of the β-aggregation propensity from 330 K constant temperature MD trajectories of trimeric seven-residue peptide systems are shown by filled circles. The segment identification number corresponds to the position of the central residue in full-length Ure2p (Table S4 in Supplementary Data). Blue triangles indicate β-aggregation propensities of single and double-point N-to-S mutants at positions 47, 48 and 49 of the stretch Ure2p$_{44–50}$ (segment 47). Open circles connected by a broken line highlight the mutational effect of the double mutation N4748S on the aggregation properties of Ure2p$_{1–94}$. Bottom: assembly kinetics of wild-type (Ure2p$_{1–94}$, red circles) and Ure2p$_{1–94}$ N4748S variant (blue triangles) monitored by ThT binding. The N4748S substitutions have a dramatic effect on the lag phase preceding assembly and validate the simulation results. Ure2p$_{1–94}$ fibrils are 4 nm wide and are shown in the electron micrograph (inset; the scale bar represent 100 nm).

aggregation process of the N-terminal domain as predicted by the implicit solvent simulations.

Concluding Discussion

A "divide-and-conquer" approach to investigate the aggregation properties of amyloid polypeptides is presented. The amino acid sequence is first decomposed in overlapping segments. Then, implicit solvent MD simulations of oligomeric systems are performed for each segment. The use of an implicit model of the solvent[36] allows for equilibrium sampling (total simulation time of hundreds of microseconds) starting from peptides well separated in space, i.e. without intermolecular contacts. To validate the structural stability of the ordered aggregates observed in the implicit solvent runs, 50 ns control simulations with explicit water are carried out for a subset of the segments.

The MD procedure is used here to determine the position dependence of the β-aggregation propensity along the polypeptide sequence, i.e. the β-aggregation profile. Despite higher computational demand with respect to analytical models recently developed to predict β-aggregation propensities,[31–33] the present method provides a structural interpretation of the β-aggregation profile, which is essential to rationalize the sequence dependence and predict mutational effects on amyloid aggregation. The use of segments to investigate the β-aggregation properties of a full-length sequence is justified, especially for parallel in-register aggregates. The dissected stretches are N-acetylated and C-amidated to reproduce their original context in the full-length sequence. Assuming in-register parallel arrangements, aggregation MD simulations of short stretches are a good approximation of the fibrillar environment and the observations made on the stretches can be extended to full-length polypeptides.

Up to now, the details of the amyloid structure and the extent to which it is uniquely defined are unclear. Initially, structural models with antiparallel β-sheets were favored.[56,57] However, solid state NMR measurements revealed that amyloid fibrils formed by $A\beta_{10-35}$[58] and by full-length $A\beta_{40}$[59,60] contain parallel β-strands exactly in register. Moreover, electron paramagnetic resonance (EPR) studies[26] and very recent vibrational dipolar coupling measurements[61] on fibrils formed by spin-labeled and isotope-labeled $A\beta_{40}$, respectively, have provided further evidence for the parallel in-register arrangement. Also, spin labeling experiments on $hIAPP_{1-37}$ fibrils have indicated an in-register parallel organization of the β-strands.[27] The compelling experimental[17,62–65] and computational[16,66] evidence that side-chains strongly influence amyloid aggregation suggests that a parallel organization of the strands in the fibrils should be, in general, preferred to antiparallel. By construction, in-register parallel arrangements favor the interactions of hydrophobic/aromatic side-chains. Hence, a preference for parallel aggregates is expected for polypeptide sequences with few charged residues. Short stretches with charged groups at the termini might prefer the antiparallel arrangement.[34]

The aggregation properties of the Alzheimer's amyloid-β peptide have been investigated by applying a two-residue shift: 18 seven-residue and 16 11-residue peptide segments were defined along the $A\beta_{42}$ sequence. Although the stretches of the two sets are rather diverse in both amino acid composition and length, the resulting amyloidogenicity profiles highlight the same region (from Val12 to Asp22) as the major hot-spot. Interestingly, this central zone is also the most prone to bind the radiolabeled full-length $A\beta_{40}$ peptide, as quantified by densitometry.[49] Since the autoradiography experiments were carried out with short (ten-residue) overlapping stretches, the comparison between *in vitro* and *in silico* results is appropriate and the former validates the latter. Furthermore, the central region includes the $K_{16}LVFF_{20}$ pentapeptide that was shown to be essential for amyloid fibril formation.[49] Although with a lower tendency, the C-terminal segment (residues 31–37) is also found to be aggregation-prone. In agreement with this *in silico* result, ThT fluorescence assays have shown that residues 30–35 (AIIGLM) promote the self-assembly by accelerating the aggregation process.[67] The enhanced β-aggregation propensity detected at the N terminus by the implicit solvent runs is rather surprising and in disagreement with solid state NMR[23] and EPR measurements.[26] It is likely that the approximations inherent to the solvation model and, in particular, the neutralization of formal charges, are too crude to correctly reproduce the behavior of polypeptide segments with many charged side-chains. Explicit solvent simulations started from parallel β-sheet conformations of segments located at the N terminus unveiled their marginal structural stability, in agreement with experiments. The structural details emerging from the simulations are consistent with the model for $A\beta_{40}$ fibrils derived from solid state NMR spectroscopy.[23] In the NMR model, the amyloid-β peptide bends to generate double-layered sheets that can pack in a parallel arrangement. Interestingly, the segments 12–24 and 31–37 correspond to the β-strands of the NMR model.

Thanks to the atomic detail provided by the MD simulations, the β-aggregation profile could be structurally characterized. Secondary structure analysis of the MD trajectories unveiled the presence of four turn-like sites along the amyloid-β sequence: S_8G_9, $G_{25}S_{26}$, $G_{29}A_{30}$, and $G_{38}V_{39}$. Interestingly, the location of the first three turns had been suggested by solution NMR,[68] solid state NMR[23] and proline scanning mutagenesis,[30] respectively. Although the four sites with turn propensity could have been detected by algorithms for secondary structure prediction, the consequences of such propensity within the context of an oligomeric system can be determined only by the

MD-simulation approach. The identified turn-like segments correspond to large drops in β-strand propensity and are located at the borders of aggregation-prone regions (Figure 3, top). Hence, their specific position on the sequence determines the location and width of the aggregation hot-spots, which are supposed to drive amyloid fibril formation and to have an influence on the fibrillar conformation of $A\beta_{42}$. In this regard, it is worth noting that although the structural model proposed by Tycko and co-workers[23] has a single turn located at residues 25–26, the original solid state NMR chemical shifts are fully compatible with the presence of a second turn located at residues 29–30, as suggested by the MD simulation results. An alternative structural model including a second turn and a different intramolecular register between the strands cannot be excluded because of the lack of experimental data in that region of the sequence and the usage of a minimization protocol by Petkova *et al.*[23] unable to investigate the whole conformational space compatible with experimental constraints.

The simulation-based approach has been further tested on the human amylin polypeptide ($hIAPP_{1–37}$) and the N-terminal domain of the yeast prion protein Ure2 ($Ure2p_{20–70}$). Unlike $A\beta_{42}$, these two polypeptide sequences contain very few charged side-chains (i.e. 2/37 and 4/51 in $hIAPP_{1–37}$ and $Ure2p_{20–70}$, respectively). Moreover, these charges are separated along the sequence (Lys1 and Arg11 in $hIAPP_{1–37}$; Arg24, Asp31, Glu38, and Arg65 in $Ure2p_{20–70}$) so that each blocked heptapeptide contains a maximum of one charge. Hence, explicit water runs were not deemed necessary. Two aggregation-prone regions have been identified along the $hIAPP_{1–37}$ sequence: a major hot-spot from Gln10 to Asn22, and a minor one from Ser28 to Val32. In contrast with experimental data obtained without blocking groups at the peptide termini,[69] the present simulation analysis indicates that the blocked NNFGAIL segment does not show high β-aggregation propensity. Interestingly, recent explicit water simulations of an octameric system of blocked NFGAIL peptides have also reported low aggregation tendency.[70] In fact, despite the usage of *ad hoc* conformational restraints, i.e. the main chains were completely restrained to ideal β-sheet conformations, only 8% of the sampled octamers were well ordered. The apparent disagreement between the simulation results (this work and work done by Wu *et al.*[70]) and the experiments suggests that short peptide stretches may show different aggregation properties if unblocked. Hence, it is more appropriate to consider blocked peptides to infer the aggregation properties of a polypeptide sequence from its segments.

The structural characterization of the β-aggregation profile of $hIAPP_{1–37}$ unveils the presence of two specific turn-like segments ($G_{24}A_{25}$ and $G_{33}S_{34}$). Similarly to what was found for $A\beta_{42}$, these two sites determine the overall shape of the aggregation profile, i.e., the location and extension of the hot-spots. It is worth noting that the $A\beta_{42}$ and $hIAPP_{1–37}$ aggregation profiles are strikingly similar (Figures 1 and 2). Both share a major aggregation hot-spot in the middle of the sequence, a less aggregation-prone region in the hydrophobic tail and a turn-like segment between them. Despite the rather low level of sequence identity (~21%) and similarity (~36%), the β-aggregation profiles suggest that $A\beta_{42}$ and $hIAPP_{1–37}$ might have similar fibrillar structures. In accord with these considerations, it has been found that $A\beta_{42}$ fibrils can act as efficient seeds for $hIAPP_{1–37}$ aggregation,[71] thus implying that at least under certain conditions $hIAPP_{1–37}$ can adopt a structure similar to that of $A\beta_{42}$ in the fibrillar form.

Taken together, the MD-simulation results of $A\beta_{42}$ and $hIAPP_{1–37}$ provide further evidence that alternating β-strands and turn (or bend) segments might be a general feature of amyloid polypeptides, as suggested by Kajava *et al.*[72,73] Assemblies with completely elongated peptide backbones are likely to be less favorable than partially folded arrangements of β-strands because of entropic effects as well as the tighter packing and minimal solvent exposure of hydrophobic residues in the latter.[23] In this view, the identified turn-like segments are expected to be very sensible to mutation and therefore optimal targets for reducing amyloid propensity.

When removed from its natural environment, $Ure2p_{1–94}$ assembles into 4 nm wide fibrils of amyloid nature (Figure 4). Three aggregation-prone regions have been identified along the sequence of $Ure2p_{20–70}$: residues 33–39, 45–49, and 55–61. Again, the presence of aggregation hot-spots intercalated with non-aggregation-prone segments has been observed. The simulation-based approach has been successfully applied to guide site-directed mutagenesis for reducing the amyloidogenic tendency of $Ure2p_{1–94}$. The double-point mutation N4748S designed *in silico* to reduce aggregation propensity has been verified experimentally (Figure 4).

In conclusion, the MD simulation approach yields the amyloidogenicity profile along a polypeptide sequence and the secondary structure propensity of its overlapping segments in the context of an ordered aggregate. The combination of both types of information is very helpful for the understanding of the sequence and structure determinants of amyloid fibril formation. The computational strategy may be ultimately used to guide the rational design of synthetic peptidic and non-peptidic molecules that hinder or prevent amyloid aggregation.

Materials and Methods

Implicit solvent simulations of aggregation

To simulate peptide aggregation, a strategy similar to that described by Gsponer *et al.*[16] was followed. Implicit

solvent simulations of aggregation were performed with the program CHARMM.[74] The oligomeric peptide systems were modeled by explicitly considering all heavy atoms and the hydrogen atoms bound to nitrogen or oxygen atoms (PARAM19 potential function[74,75]). The remaining hydrogen atoms are considered as part of the carbon atoms to which they are covalently bound (extended atom approximation). An implicit model based on the solvent-accessible surface was used to describe the main effects of the aqueous solvent on the solute.[36] The CHARMM PARAM19 default cutoffs for long range interactions were used, i.e., a shift function[74] was employed with a cutoff at 7.5 Å for both the electrostatic and van der Waals terms. The model is not biased toward any particular secondary structure type. In fact, the same force field and implicit solvent model have been recently used in MD simulations of aggregation,[16,48] folding of structured peptides (α-helices and β-sheets) ranging in size from 15 to 31 residues,[76–78] and small proteins of about 60 residues.[79,80] Moreover, the in-register parallel packing of three GNNQQNY peptides predicted by this model[16] has been recently validated by the X-ray microcrystal structure of the cross-β spine:[18] the β-strand alignment, the stacking interactions of the tyrosine rings and the hydrogen bonds between amide groups are essentially identical (compare Figure 2 of Gsponer *et al.*[16] with Figure 2(e) of Nelson *et al.*[18]). The same force field, solvation model and simulation protocol have been applied to polypeptide segments experimentally known not to form amyloid structures, i.e. the nonapeptide SQNGNQQRG (corresponding to Sup35 residues 17–25 with the Gln/Arg mutation at position 24, which showed solubility *in vivo* and *in vitro*[81]) and the alanine heptapeptide.[82] No ordered β-aggregate was observed in these control simulations,[16,48] which is particularly remarkable for the SQNGNQQRG sequence given its similarity to the amyloidogenic GNNQQNY peptide.

In the case of $A\beta_{42}$, the following constant temperature MD runs were performed: (i) 18 simulations of three seven-residue peptide copies at 330 K to monitor β-aggregation propensity along the sequence and highlight possible aggregation hot-spots; (ii) 18 simulations of three seven-residue peptide copies at 310 K to investigate the temperature effect on the aggregation properties of the overlapping stretches; (iii) 18 simulations of six seven-residue peptide copies at 330 K to investigate the effect of the system size, i.e. the number of molecules in the simulation box, on both β-aggregation propensity and structural properties of the aggregates; (iv) 13 310 K and four 330 K runs of three seven-residue peptide copies to predict mutational effects on the aggregation properties of $A\beta_{42}$; and (v) 16 simulations with three 11-residue peptide copies at 330 K to study the dependency of β-aggregation propensity on the length of the overlapping segments. To guarantee the correct sampling of peptide conformational space in physiological conditions,[48] the self-assembly of long segments (11-residue) was investigated by replica exchange molecular dynamics (REMD) simulations. In an REMD run, different copies of the system ("replicas") are simulated at the same time but at different temperature values. Each replica evolves independently by MD and every t_{swap} states i,j with neighbor temperatures are swapped (by velocity rescaling) with a probability $w_{ij}=\exp(-\Delta)$,[83] where $\Delta\equiv(\beta_i-\beta_j)(E_j-E_i)$, $\beta=1/kT$ and E is the effective energy (potential and solvation energy). During the simulation, each replica visits all temperatures of the set and realizes a free random walk in temperature space. High-temperature simulation segments facilitate the crossing of the energy barriers while low temperature ones explore energy minima in detail. Thus, the temperature swapping determines a random walk in energy space, which improves sampling efficiency. In this study, ten replicas were used with temperatures (in K): 294, 306, 318, 330, 343, 356, 369, 383, 397, 413. By using fixed value of $\Delta t_{swap}=10{,}000$ MD steps (20 ps),[48] temperature values were adjusted by trial and error until the acceptance ratios of exchange between neighbor temperatures converged to values between 40% and 50%.

In a similar fashion, to determine the β-aggregation profile and identify the aggregation hot-spots along both $hIAPP_{1–37}$ and $Ure2p_{20–70}$, 16 and 23 simulations of three seven-residue peptide copies at 330 K were carried out, respectively. Finally, ten simulations of three seven-residue peptide copies were performed at 330 K to investigate mutational effects on $Ure2p_{1–94}$ aggregation properties.

All implicit solvent simulations were performed starting from random conformations, positions, and orientations of the peptide copies. In the initial random positions there was no intermolecular contact, i.e. the peptides were separated in space. Each system was simulated in a cubic box whose side was adjusted to yield a sample concentration of 5 mg/ml. Langevin dynamics with a friction value of 0.15 ps^{-1} was used. This friction coefficient is much smaller than that of water (43 ps^{-1} at 330 K) to allow for sufficient sampling within the microsecond time-scale of the simulation. The small friction does not influence the thermodynamic properties of the system. The SHAKE algorithm[84] was used to fix the length of the covalent bonds involving hydrogen atoms, which allows an integration time-step of 2 fs. Furthermore, the non-bonded interactions were updated every ten dynamics steps and coordinate frames were saved every 20 ps for a total of 5×10^4 conformations/μs. On a 2.1 GHz Athlon processor, a 1 μs run requires approximately 10.4 days, 25.2 days and 22.3 days for three, seven-residue peptides, six, seven-residue peptides, and three 11-residue peptides, respectively. Simulations were run on a Beowulf cluster for a total simulation time of 0.35 ms, 0.02 ms and 0.05 ms for $A\beta_{42}$, $hIAPP_{1–37}$ and $Ure2p_{20–70}$, respectively.

Explicit solvent simulations started from ordered aggregates

For a subset of four $A\beta_{42}$ seven-residue stretches ($A\beta_{1–7}$, $A\beta_{5–11}$, $A\beta_{13–19}$ and $A\beta_{29–35}$), explicit solvent MD simulations were carried out starting from in-register parallel β-sheet conformations. The starting structure of each run was selected among the implicit solvent conformations with a fraction of parallel contacts larger than 0.85 (these contacts were defined following the procedure described by Gsponer *et al.*[16]). Explicit solvent simulations were performed with the program NAMD2[85] using the CHARMM all-hydrogen force field (PARAM22 potential function[86]) along with the TIP3P model for water molecules.[87] Non-polar hydrogen atoms were added by CHARMM (HBUILD module) and their position was optimized *in vacuo*. The resulting structure was solvated in a water box of appropriate dimensions so that the distance between periodic images was not smaller than 25 Å. Chloride and sodium ions were added to neutralize the systems, yielding a salt concentration of 150 mM. Long-range

electrostatic forces were accounted for by using the particle mesh Ewald summation method[88] with real space cutoff distance of 10 Å and a grid width of 0.93 Å. The simulations were run at constant temperature (310 K) and pressure (1 atm) by applying the Berendsen thermostat[89] with a coupling decay time of 1 ps and the hybrid Nose–Hoover Langevin pressure control.[90] The SHAKE algorithm[84] was used to allow for an integration time-step of 2 fs. Solvent molecules and counterions were equilibrated at 310 K while holding the peptide system rigid for 1 ns. Two 0.5 ns equilibration cycles were then performed applying a harmonic constraint to all peptide atoms with force constants of 1.0 and 0.1 kcal/mol Å^2, respectively. Upon releasing the constraints, 50 ns production runs were performed for each peptide system.

Order parameters and β-aggregation propensity

The nematic order parameter $\bar{P}_2$ was considered to monitor the aggregation process as described by Cecchini *et al*.[48] This order parameter is widely used for studying the properties of anisotropic fluids such as liquid crystals[91–93] and is defined as:

$$\bar{P}_2 = \frac{1}{N}\sum_{i=1}^{N}\frac{3}{2}(\hat{\mathbf{z}}_i\cdot\hat{\mathbf{d}})^2 - \frac{1}{2} \quad (1)$$

where $\hat{\mathbf{d}}$ (the director) is a unit vector defining the preferred direction of alignment, $\hat{\mathbf{z}}_i$ is a suitably defined molecular vector, and N is the number of molecules in the simulation box, i.e., three or six peptides in this study. The director is defined as the eigenvector of the ordering matrix[94] that corresponds to the largest positive eigenvalue. Here, the molecular vectors $\hat{\mathbf{z}}_i$ were defined as unit vectors linking the peptide's termini (from the N to the C terminus). The nematic $\bar{P}_2$ describes the orientational order of the system and discriminates between ordered and disordered conformations.

As it has been recently shown,[48] the average over the canonical ensemble of $\bar{P}_2$ is descriptive of the thermodynamic stability of the ordered state of oligomeric peptide systems. This scalar value, which ranges from 0 (complete disorder) to 1 (perfect order), can then be used to estimate and compare the amyloidogenic propensity of different peptide sequences. Here, the value of $\bar{P}_2$ averaged over the MD trajectory, referred to as β-aggregation propensity, is used to determine the β-aggregation profiles.

Progress variables

Radius of gyration

The radius of gyration of the oligomeric system R_g was considered to monitor the "condensation" equilibrium along the simulation trajectories. Conformations of the system producing non-interacting peptides, namely conformations where all inter-peptide atomic distances are larger than the long-range interactions cutoffs (7.5 Å in this case), were used to determine R_g^C. In other words, R_g^C is the lowest value of the radius of gyration measured for the snapshots when the three peptides are far apart from each other. Conformations with one or more isolated peptides ($R_g > R_g^C$) describe the "uncondensed state" of the system.

Secondary structure

Strings of secondary structure (SS) were considered to monitor peptide conformational changes along the trajectory. For each oligomeric snapshot (Cartesian coordinates of the atomic nuclei) the SS of each chain was calculated.[95] The resulting strings of SS elements (one element per residue) were used to describe peptide conformations and monitor the aggregation process. The SS alphabet includes four possible letters: "H", "E", "T" and "–", which stand for α-helix, extended (β-strand), β-turn or bend, and random coil, respectively. Terminal caps as well as N and C-terminal residues are always assigned a "–". Albeit devoid of the atomic detail, SS strings are useful because they provide an intuitive description of the shape of the peptide backbone. However, these strings are suitable to monitor peptide conformational changes only at a coarse-grained level. In fact, in a perfect in-register arrangement a capped hendecapeptide would present the following SS string "–EEEEEEEEE–" independently of the polarity of the assembly, i.e. the number of parallel and antiparallel β-strands.

Mutagenesis

The N-terminal domain of Ure2p (Ure2p$_{1–94}$) is highly insoluble and forms inclusion bodies in *E. coli*. In contrast, it is soluble when attached to the C-terminal domain of the protein. We therefore engineered a specific cleavage site between the two domains in order to generate soluble Ure2p$_{1–94}$ at the onset of the assembly reaction by cleavage with the specific protease Factor Xa as described by Bousset *et al.*[96] The variant Ure2p-N4748S expression vector was obtained by site-directed mutagenesis. Mutagenesis was achieved using the QuickChange site-directed mutagenesis kit (Stratagene Europe, Amsterdam, The Netherlands) and the primers 5′-CAGGTGTAAATAATAATAGTAGTAA-CAATAGCAGTAGTAATAAC-3′ and 5′-GTTATTAC-TACTGCTATTGTTACTACTATTATTATTTACACCTG-3′.

Protein purification, generation of soluble Ure2p$_{1–94}$ and assembly of Ure2p$_{1–94}$ into fibrils

Recombinant Ure2p-I91EGR94 and Ure2p-N4748S-I91EGR94 were over expressed as soluble proteins in *E. coli* and purified as described.[46] The Ure2p$_{1–94}$ fragment was generated as described by Bousset *et al.*[96] The assembly reactions were monitored using thioflavin T binding ,[97] using a Quantamaster QM 2000-4 spectrofluorimeter (Photon Technology International, Inc. NJ). Ure2p$_{1–94}$ fibrils were also examined following negative staining with 1% uranyl acetate on carbon-coated grids (200-mesh) in a Philips EM 410 electron microscope (Philips Inc., The Netherlands).

Acknowledgements

We thank R. Pellarin for introducing periodic boundary conditions in the SASA module in CHARMM (version 29) and S. Dubois for excellent technical assistance. We are grateful to E. Guarnera, F. Rao and G. G. Tartaglia for helpful discussions, and to Professor F. E. Cohen for suggesting the control runs with explicit solvent. We thank

A. Widmer (Novartis Pharma, Basel) for providing the molecular modeling program Wit!P, which was used for visual analysis of the trajectories. We are very grateful to Dr M. Seeber (University of Modena and Reggio Emilia) for his computer program for the efficient analysis of CHARMM trajectories and to Dr M. Schaefer (Sygenta, Basel, Switzerland) for providing the program used for the clustering with RMSD between all pairs of structures. The simulations were performed on the Matterhorn Beowulf cluster at the Computing Center of the University of Zurich. We thank C. Bolliger and Dr A. Godknecht for setting up the cluster and the Canton of Zurich for generous hardware support. This work was supported by grants from the Swiss National Competence Center in Neural Plasticity and Repair (NCCR) and the Swiss National Science Foundation to AC and the French Ministry of Education and Research through the GIS Prions grant to RM.

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2006.01.009

References

1. Thomas, P., Qu, B. & Pedersen, P. (1995). Defective protein folding as a basis of human disease. *Trends Biochem. Sci.* **20**, 456–459.
2. Dobson, C. M. (2001). The structural basis of protein folding and its links with human disease. *Phil. Trans. Roy. Soc. Ser. B*, **356**, 133–145.
3. Horwich, A. (2002). Protein aggregation in disease: a role for folding intermediates forming specific multimeric interactions. *J. Clin. Invest.* **110**, 1221–1232.
4. Westermark, P., Benson, M., Buxbaum, J., Cohen, A., Frangione, B., Ikeda, S. *et al*. (2002). Amyloid fibril protein nomenclature, *Amyloid*, **9**, 197–200.
5. Dobson, C. M. (1999). Protein misfolding, evolution and disease. *Trends Biochem. Sci.* **24**, 329–332.
6. Perutz, M. F. (1999). Glutamine repeats and neurodegenerative diseases: molecular aspects. *Trends Biochem. Sci.* **24**, 58–63.
7. Blake, C. & Serpell, L. (1996). Synchrotron X-ray studies suggest that the core of the transthyretin amyloid fibril is a continuous β-sheet helix. *Structure*, **4**, 989–998.
8. Malinchik, S. B., Inouye, H., Szumowski, K. E. & Kirschner, D. A. (1998). Structural analysis of Alzheimer's β_{1-40} amyloid: protofilament assembly of tubular fibrils. *Biophys. J.* **74**, 537–545.
9. Sunde, M. & Blake, C. C. F. (1997). The structure of amyloid fibrils by electron microscopy and X-ray diffraction. *Advan. Protein Chem.* **50**, 123–159.
10. Guijarro, J., Sunde, M., Jones, J., Campbell, I. & Dobson, C. (1998). Amyloid fibril formation by an SH3 domain. *Proc. Natl Acad. Sci. USA.* **95**, 4224–4228.
11. Konno, T., Murata, K. & Nagayama, K. (1998). Amyloid-like aggregates of a plant protein: a case of a sweet-tasting protein, monellin. *FEBS Letters*, **95**, 4224–4228.
12. Fandrich, M., Fletcher, M. & Dobson, C. (2001). Amyloid fibrils from muscle myoglobin. *Nature*, **410**, 165–166.
13. Dobson, C. M. (2003). Protein folding and misfolding. *Nature*, **426**, 884–890.
14. Gazit, E. (2002). A possible role for π-stacking in the self-assembly of amyloid fibrils. *FASEB J.* **16**, 77–83.
15. Tartaglia, G. G., Pellarin, R., Cavalli, A. & Caflisch, A. (2004). The role of aromaticity, exposed surface and dipole moment in determining protein aggregation rates. *Protein Sci.* **13**, 1939–1941.
16. Gsponer, J., Habertür, U. & Caflisch, A. (2003). The role of side-chain interactions in the early steps of aggregation: molecular dynamics simulations of an amyloid-forming peptide from the yeast prion Sup35. *Proc. Natl Acad. Sci. USA*, **100**, 5154–5159.
17. Lopez de la Paz, M. & Serrano, L. (2004). Sequence determinants of amyloid fibril formation. *Proc. Natl Acad. Sci. USA.* **101**, 87–92.
18. Nelson, R., Sawaya, M., Balbirnie, M., Madsen, A., Riekel, C., Grothe, R. & Eisenberg, D. (2005). Structure of the cross-β spine of amyloid-like fibrils. *Nature*, **435**, 773–778.
19. Griffiths, J., Ashburn, T., Auger, M., Costa, P., Griffin, R. & Lansbury, P. (1995). Rotational resonance solid-state NMR elucidates a structural model of pancreatic amyloid. *J. Am. Chem. Soc.* **117**, 3539–3546.
20. Burkoth, T., Benzinger, T., Urban, V., Morgan, D., Gregory, D., Thiyagarajan, P. *et al*. (2000). Structure of the β-amyloid (10–35). *J. Am. Chem. Soc.* **122**, 7883–7889.
21. Benzinger, T., Gregory, D., Burkoth, T., Miller-Auer, H., Lynn, D., Botto, R. & Meredith, S. (2000). Two-dimensional structure of β-amyloid (10–35) fibrils. *Biochemistry*, **39**, 3491–3499.
22. Antzutkin, O., Leapman, R., Balbach, J. & Tycko, R. (2002). Supramolecular structural constraints on Alzheimer's β-amyloid fibrils from electron microscopy and solid-state nuclear magnetic resonance. *Biochemistry*, **41**, 15436–15450.
23. Petkova, A. T., Ishii, Y., Balbach, J. J., Antzutkin, O. N., Leapman, R. D., Delaglio, F. & Tycko, R. (2002). A structural model for Alzheimer's β-amyloid fibrils based on experimental constraints from solid state NMR. *Proc. Natl Acad. Sci. USA.* **99**, 16742–16747.
24. Jaroniec, C., MacPhee, C., Bajaj, V., McMahon, M., Dobson, C. & Griffin, R. (2004). High-resolution molecular structure of a peptide in an amyloid fibril determined by magic angle spinning NMR spectroscopy. *Proc. Natl Acad. Sci. USA.* **101**, 711–716.
25. Serag, A., Altenbach, C., Gingery, M., Hubbell, W. & Yeates, T. (2001). Identification of a subunit interface in transthyretin amyloid fibrils: evidence for self-assembly from oligomeric building blocks. *Biochemistry*, **40**, 9089–9096.
26. Torok, M., Milton, S., Kayed, R., Wu, P., McIntire, T., Glabe, C. & Langen, R. (2002). Structural and dynamic features of Alzheimer's Aβ peptide in amyloid fibrils studied by site-directed spin labeling. *J. Biol. Chem.* **277**, 40810–40815.
27. Jayasinghe, S. & Langen, R. (2004). Identifying structural features of fibrillar islet amyloid polypep-

tide using site-directed spin labeling. *J. Biol. Chem.* **279**, 48420–48425.
28. Jimenez, J., Guijarro, J., Orlova, E., Zurdo, J., Dobson, C., Sunde, M. & Saibil, H. (1999). Cryo-electron microscopy structure of an SH3 amyloid fibril and model of the molecular packing. *EMBO J.* **18**, 815–821.
29. Kishimoto, A., Hasegawa, K., Suzuki, H., Taguchi, H., Namba, K. & Yoshida, M. (2004). β-helix is a likely core structure of yeast prion Sup35 amyloid fiber-scryo-electron microscopy structure of an SH3 amyloid. *Biochim. Biophys. Acta*, **315**, 739–745.
30. Williams, A. D., Portelius, E., Kheterpal, I., Guo, J., Cook, K. D., Xu, Y. & Wetzel, R. (2004). Mapping Aβ amyloid fibril secondary structure using scanning proline mutagenesis. *J. Mol. Biol.* **335**, 833–842.
31. Dubay, K. F., Pawar, A. P., Chiti, F., Zurdo, J., Dobson, C. M. & Vendruscolo, M. (2004). Prediction of the absolute aggregation rates of amyloidogenic polypeptide chains. *J. Mol. Biol.* **341**, 1317–1326.
32. Fernandez-Escamilla, A. M., Rousseau, F., Schymkowitz, J. & Serrano, L. (2004). Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nature Biotechnol.* **22**, 1302–1306.
33. Tartaglia, G. G., Pellarin, R., Cavalli, A. & Caflisch, A. (2005). Prediction of aggregation rate and aggregation-prone segments in polypeptide chains. *Protein Sci.* **14**, 2723–2734.
34. Lopez de la Paz, M., de Mori, G., Serrano, L. & Colombo, G. (2005). Sequence dependence of amyloid fibril formation: insights from molecular dynamics simulations. *J. Mol. Biol.* **349**, 583–596.
35. Buchete, N., Tycko, R. & Hummer, G. (2005). Molecular dynamics simulations of Alzheimer's β-amyloid protofilaments. *J. Mol. Biol.* **353**, 804–821.
36. Ferrara, P., Apostolakis, J. & Caflisch, A. (2002). Evaluation of a fast implicit solvent model for molecular dynamics simulations. *Proteins: Struct. Funct. Genet.* **46**, 24–33.
37. Selkoe, D. J. (1999). Translating cell biology into therapeutic advances in Alzheimer's disease. *Nature*, **399**, A23–A31.
38. Citron, M. (2004). β-Secretase inhibition for the treatment of Alzheimer's disease: promise and challenge. *Trends Pharmacol. Sci.* **25**, 92–97.
39. Westermark, P. & Wilander, E. (1978). Influence of amyloid deposits on islet volume in maturity onset diabetes-mellitus. *Diabetologia*, **15**, 417–421.
40. King, H., Aubert, R. & Herman, W. (1998). Global burden of diabetes, 1995–2025–prevalence, numerical estimates, and projections. *Diabetes Care*, **21**, 1414–1431.
41. Makin, S. & Serpell, L. C. (2004). Structural characterization of islet amyloid polypeptide fibrils. *J. Mol. Biol.* **335**, 1279–1288.
42. Wickner, R. (1994). [URE3] as an altered Ure2 protein–evidence for a prion analog in *Saccharomyces cerevisiae*. *Science*, **264**, 566–569.
43. Courchesne, W. & Magasanik, B. (1988). Regulation of nitrogen assimilation in *Saccharomyces cerevisiae*–roles of the ure2 and gln3 genes. *J. Bacteriol.* **170**, 708–713.
44. Lacroute, F. (1971). Non-mendelian mutation allowing ureidosuccinic acid uptake in yeast. *J. Bacteriol.* **106**, 519–522.
45. Masison, D., Maddelein, M. & Wickner, R. (1997). The prion model for [URE3] of yeast: spontaneous generation and requirements for propagation. Prion-inducing domain of yeast Ure2p and protease. *Proc. Natl Acad. Sci. USA.* **94**, 12503–12508.
46. Thual, C., Komar, A., Bousset, L., Fernandez-Bellot, E., Cullin, C. & Melki, R. (1999). Structural characterization of *Saccharomyces cerevisiae* prion-like protein Ure2. *J. Biol. Chem.* **274**, 13666–13674.
47. Taylor, K., Cheng, N., Williams, R., Steven, A. & Wickner, R. (1999). Prion domain initiation of amyloid formation *in vitro* from native Ure2p. *Science*, **283**, 1339–1343.
48. Cecchini, M., Rao, F., Seeber, M. & Caflisch, A. (2004). Replica exchange molecular dynamics simulations of amyloid peptide aggregation. *J. Chem. Phys.* **121**, 10748–10756.
49. Tjernberg, L. O., Näslund, J., Lindqvist, F., Johansson, J., Karlstrom, A. R., Thyberg, J. *et al*. (1996). Arrest of β-amyloid fibril formation by a pentapeptide ligand. *J. Biol. Chem.* **271**, 8545–8548.
50. Riek, R., Güntert, P., Döbeli, H., Wipf, B. & Wüthrich, K. (2001). NMR studies in aqueous solution fail to identify significant conformational differences between the monomeric forms of two Alzheimer peptides with widely different plaque-competence, $A\beta(1–40)^{ox}$ and $A\beta(1–42)^{ox}$. *Eur. J. Biochem.* **268**, 5930–5936.
51. Wurth, C., Guimard, N. & Hecht, M. (2002). Mutations that reduce aggregation of the Alzheimer's $A\beta_{42}$ peptide: an unbiased search for the sequence determinants of Aβ amyloidogenesis. *J. Mol. Biol.* **319**, 1279–1290.
52. Nilsberth, C., Westlind-Danielsson, A., Eckman, C., Condron, M., Axelman, K., Forsell, C. *et al*. (2001). The "Arctic" APP mutation (E693G) causes Alzheimer's disease by enhanced Aβ protofibril formation. *Nature Neurosci.* **4**, 887–893.
53. Päiviö, A., Jarvet, J., Gräslund, A., Lannfelt, L. & Westlind-Danielsson, A. (2004). Unique physicochemical profile of β-amyloid peptide variant $A\beta_{1–40}$ E22G protofibrils: conceivable neuropathogen in arctic mutant carriers. *J. Mol. Biol.* **339**, 145–159.
54. Massi, F., Klimov, D., Thirumalai, D. & Straub, J. E. (2002). Charge states rather than propensity for β-structure determine enhanced fibrillogenesis in wild-type Alzheimer's β-amyloid peptide compared to E22Q Dutch mutant. *Protein Sci.* **11**, 1639–1647.
55. Mazor, Y., Gilead, S., Benhar, I. & Gazit, E. (2002). Identification and characterization of a novel molecular-recognition and self-assembly domain within the islet amyloid polypeptide. *J. Mol. Biol.* **322**, 1013–1024.
56. Tjernberg, L. O., Callaway, D. J. E., Tjernberg, A., Hahne, S., Lilliehöök, C., Terenius, L. *et al*. (1999). A molecular model of Alzheimer amyloid β-peptide fibril formation. *J. Biol. Chem.* **274**, 12619–12625.
57. Li, L., Darden, T., Bartolotti, L., Kominos, D. & Pedersen, L. (1999). An atomic model for the pleated β-sheet structure of Aβ amyloid protofilaments. *Biophys. J.* **76**, 2871–2878.
58. Benzinger, T., Gregory, D., Burkoth, T., Miller-Auer, H., Lynn, D., Botto, R. & Meredith, S. (1998). Propagating structure of Alzheimer's β-amyloid (10–35) is parallel β-sheet with residues in exact register. *Proc. Natl Acad. Sci. USA*, **95**, 13407–13412.
59. Antzutkin, J., Balbach, O. N., Leapman, R., Rizzo, N., Reed, J. & Tycko, R. (2000). Multiple quantum solid-state NMR indicates a parallel, not antiparallel, organization of β-sheets in Alzheimer's β-amyloid fibrils. *Proc. Natl. Acad. Sci. USA.* **97**, 13045–13050.
60. Balbach, J., Petkova, A., Oyler, N., Antzutkin, O., Gordon, D., Meredith, S. & Tycko, R. (2002). Supramolecular structure in full-length Alzheimer's

β-amyloid fibrils: evidence for a parallel β-sheet organization from solid-state nuclear magnetic resonance. *Biophys. J.* **83**, 1205–1216.
61. Paul, C. & Axelsen, P. (2005). β sheet structure in amyloid-β fibrils and vibrational dipolar coupling. *J. Am. Chem. Soc.* **127**, 5754–5755.
62. West, M., Wang, W., Patterson, J., Mancias, J., Beasley, J. & Hecht, M. (1999). *De novo* amyloid proteins from designed combinatorial libraries. *Proc. Natl Acad. Sci. USA*, **96**, 11211–11216.
63. Lopez de la Paz, M., Goldie, K., Zurdo, J., Lacroix, E., Dobson, C. M., Hoenger, A. & Serrano, L. (2002). *De novo* designed peptide-based amyloid fibrils. *Proc. Natl Acad. Sci. USA*, **99**, 16052–16057.
64. Hammarstrom, P., Jiang, X., Hurshman, A., Powers, E. & Kelly, J. (2002). Sequence-dependent denaturation energetics: a major determinant in amyloid disease diversity. *Proc. Natl Acad. Sci. USA*, **99**, 16427–16432.
65. Chiti, F., Stefani, M., Taddei, N., Ramponi, G. & Dobson, C. (2003). Rationalization of the effects of mutations on peptide and protein aggregation rates. *Nature*, **424**, 805–808.
66. Klimov, D. & Thirumalai, D. (2003). Dissecting the assembly of $A\beta_{16-22}$ amyloid peptides into antiparallel β sheets. *Structure*, **11**, 295–307.
67. Liu, R., McAllister, C., Lyubchenko, Y. & Sierks, M. R. (2003). Residues 17–20 and 30–35 of β-amyloid play crucial roles in aggregation. *J. Neurosci. Res.* **75**, 162–171.
68. Hou, L., Shao, H., Zhang, Y., Hua, L., Menon, N. K., Neuhaus, E. *et al*. (2004). Solution NMR studies of the Aβ(1–40) and Aβ(1–42) peptides established that Met35 oxidation state affects the mechanism of amyloid formation. *J. Am. Chem. Soc.* **126**, 1992–2005.
69. Tenidis, K., Waldner, M., Bernhagen, J., Fischle, W., Bergmann, M., Weber, M., Kapurniotu, A. *et al*. (2000). Identification of a penta- and hexapeptide of islet amyloid polypeptide (IAPP) with amyloidogenic and cytotoxic properties. *J. Mol. Biol.* **295**, 1055–1071.
70. Wu, C., Lei, H. & Duan, Y. (2005). The role of phe in the formation of well-ordered oligomers of amyloidogenic hexapeptide (NFGAIL) observed in molecular dynamics simulations with explicit solvent. *Biophys. J.* **88**, 2897–2906.
71. O'Nuallain, B., Williams, A., Westermark, P. & Wetzel, R. (2004). Seeding specificity in amyloid growth induced by heterologous fibrils. *J. Biol. Chem.* **279**, 17490–17499.
72. Kajava, A., Baxa, U., Wickner, R. & Steven, A. (2004). A model for Ure2p prion filaments and other amyloids: the parallel superpleated β-structure. *Proc. Natl Acad. Sci. USA*, **101**, 7885–7890.
73. Kajava, A., Ueli, A. & Steven, A. (2005). The parallel superpleated beta-structure as a model for amyloid fibrils of human amylin. *J. Mol. Biol.* **348**, 247–252.
74. Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S. & Karplus, M. (1983). CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **4**, 187–217.
75. Neria, E., Fischer, S. & Karplus, M. (1996). Simulation of activation free energies in molecular systems. *J. Chem. Phys.* **105**, 1902–1921.
76. Hiltpold, A., Ferrara, P., Gsponer, J. & Caflisch, A. (2000). Free energy surface of the helical peptide $Y(MEARA)^6$. *J. Phys. Chem. B*, **104**, 10080–10086.
77. Ferrara, P. & Caflisch, A. (2000). Folding simulations of a three-stranded antiparallel β-sheet peptide. *Proc. Natl Acad. Sci. USA*, **97**, 10780–10785.
78. Ferrara, P. & Caflisch, A. (2001). Native topology or specific interactions: what is more important for peptide folding? *J. Mol. Biol.* **306**, 837–850.
79. Gsponer, J. & Caflisch, A. (2001). Role of native topology investigated by multiple unfolding simulations of four SH3 domains. *J. Mol. Biol.* **309**, 285–298.
80. Gsponer, J. & Caflisch, A. (2002). Molecular dynamics simulations of protein folding from the transition state. *Proc. Natl Acad. Sci. USA*, **99**, 6719–6724.
81. Balbirnie, M., Grothe, R. & Eisenberg, D. (2001). An amyloid-forming peptide from the yeast prion Sup35 reveals a dehydrated β-sheet structure for amyloid. *Proc. Natl Acad. Sci. USA*, **98**, 2375–2380.
82. Perutz, M. F., Pope, B. J., Owen, D., Wanker, E. E. & Scherzinger, E. (2002). Aggregation of proteins with expanded glutamine and alanine repeats of the glutamine-rich and asparagine-rich domains of Sup35 and of the amyloid β-peptide of amyloid plaques. *Proc. Natl Acad. Sci. USA*, **99**, 5596–5600.
83. Sugita, Y. & Okamoto, Y. (1999). Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Letters*, **314**, 141–151.
84. Ryckaert, J. P., Ciccotti, G. & Berendsen, H. J. C. (1977). Numerical integration of the Cartesian equation of motion of a system with constraints: molecular dynamics of *n*-alkanes. *J. Comp. Phys.* **23**, 327–341.
85. Kalè, L., Skeel, R., Bhandarkar, M., Brunner, R., Gursoy, A., Krawetz, N. *et al*. (1999). Namd2: greater scalability for parallel molecular dynamics. *J. Comput. Phys.* **151**, 283–312.
86. MacKerell, A., Jr, Bashford, D., Bellott, M., Dunbrack, R., Jr, Evanseck, J., Field, M. *et al*. (1998). All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B*, **102**, 3586–3616.
87. Jorgensen, W. L., Chandrasekhar, J., Madura, J., Impey, R. W. & Klein, M. L. (1983). Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926–935.
88. Darden, T., York, D. & Pedersen, L. (1993). Particle mesh Ewald–an $N\log(N)$ method for ewald sums in large systems. *J. Chem. Phys.* **98**, 10089–10092.
89. Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A. & Haak, J. R. (1984). Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **81**, 3684–3690.
90. Hoover, W. (1985). Canonical dynamics: equilibrium phase-space distributions. *Phys. Rev. A*, **31**, 1695–1697.
91. Chandrasekhar, S. (1992). *Liquid Crystals*, Cambridge University Press, Cambridge, England.
92. de Gennes, P. G. & Prost, J. (1993). *The Physics of Liquid Crystals* 2nd edit., Oxford University Press, Oxford.
93. Zannoni, C. (2001). Molecular design and computer simulations of novel mesophases. *J. Mater. Chem.* **11**, 2637–2646.
94. Allen, M. P. & Tildesley, D. J. (1987). *Computer Simulation of Liquids*, Oxford Science Publications, Oxford, UK.
95. Andersen, C. A. F., Palmer, A. G., Brunak, S. & Rost, B. (2002). Continuum secondary structure captures protein flexibility. *Structure*, **10**, 174–184.

96. Bousset, L., Redeker, V., Decottignies, P., Dubois, S., LeMarechal, P. & Melki, R. (2004). Structural characterization of the fibrillar form of the yeast *Saccharomyces cerevisiae* prion Ure2p. *Biochemistry,* **43**, 5022–5032.

97. McParland, V., Kad, N., Kalverda, A., Brown, A., Kirwin-Jones, P., Hunter, M. *et al*. (2000). Partially unfolded states of β_2-microglobulin and amyloid formation *in vitro*. *Biochemistry,* **39**, 8735–8746.

Edited by F. E. Cohen

(Received 11 July 2005; received in revised form 21 November 2005; accepted 4 January 2006)
Available online 26 January 2006

Supplementary Material

A Molecular Dynamics Approach to the Structural Characterization of Amyloid Aggregation

M. Cecchini, R. Curcio, M. Pappalardo[†], R. Melki[*], and A. Caflisch[‡]

Department of Biochemistry,

University of Zurich,

Winterthurerstrasse 190,

CH-8057 Zurich, Switzerland

tel: +41 44 635 55 21,

fax: +41 44 635 68 62,

[‡]*:Corresponding author,*

e-mail: caflisch@bioc.unizh.ch

[†]*Dipartimento di Scienze Chimiche,*

Università di Catania,

Viale Andrea Doria 6,

95125 Catania, Italy

[*]*Laboratoire d'Enzymologie et Biochimie Structurales, CNRS,*

Avenue de la Terrasse,

91198 Gif-sur-Yvette, France

(Dated: March 13, 2006)

1. Single-peptide cluster analysis

A single-peptide cluster analysis was performed to investigate the conformations visited by each peptide segment in the perturbation field of its oligomeric system and to determine the statistical relevance of the conformations sampled. The simulation trajectory of the oligomeric system was first split into n single-peptide trajectories, where n is the number of peptide copies in the simulation box. For each snapshot, the Cartesian coordinates of single peptides were extracted and collected according to their identification number, i.e., coordinates of copy 1 were gathered in the first trajectory, coordinates of copy 2 in the second and so on. Single-peptide trajectories were then merged one after the other. The resulting trajectory, which is made of a number of snapshots n times larger than the original trajectory, underwent a cluster analysis based on structural similarity[1]. Single-peptide conformations were grouped according to the C_α–C_β root-mean-square deviation (RMSD) for *each* pairs of structures after optimal superposition[2]. A 1.5 Å RMSD cutoff was used to determine the cluster centers. The clustering algorithm proceeded iteratively until each structure was assigned to a cluster center.

2. Role of condensation in amyloid peptide aggregation

To further investigate the role of the condensation equilibrium in amyloid peptide aggregation, the simulation trajectories of two non aggregation-prone stretches ($A\beta_{25-31}$ and $A\beta_{29-35}$) were analyzed in detail. Unlike the free-energy profiles along $\overline{P_2}$ and R_g, a single-peptide cluster analysis (see above) revealed that the two oligomeric systems have intrinsically different β-aggregation propensities. For both systems, the most populated clusters sampled along the whole trajectory are shown in Fig. S1 (top). For $A\beta_{29-35}$, already the third most populated cluster (4.1%) is the extended conformation which favors β-sheet formation. Moreover, if one considers only the fraction of condensed conformations sampled along the trajectory ($R_g < {R_g}^C$) the statistical weight of the cluster increases substantially and the extended conformation becomes the most populated cluster (6.3%). Hence, $A\beta_{29-35}$ does have a certain β-sheet propensity, which due to the condensation equilibrium remains hidden at 330 K and cannot be detected by the order parameter analysis. For $A\beta_{25-31}$, the situation is different: the cluster producing the extended conformation ranks 19^{th} (statistical

weight of only 0.7%). However, even considering the condensed fraction, its ranking does not improve. Irrespective of the degree of condensation $A\beta_{25-31}$ does not show any β-sheet propensity. Thus, the condensation propensity, though not sufficient to describe amyloidogenicity, is a necessary condition to promote the growth of ordered amyloid nuclei. It is not sufficient because amyloidogenic sequences must also contain β-sheet propensity.

3. Effect of the number of peptides on β-aggregation

To investigate the effect of the dimensionality of the system on the $A\beta_{42}$ β-aggregation profile, the self-assembly process of hexamers was considered. Eighteen MD simulations of six peptides were performed (see Table I) and each trajectory was analyzed by means of the order parameters. At 330 K, β-aggregation propensities ranged between 0.28 and 0.71, thus confirming a strong heterogeneity in aggregation tendencies. A direct comparison with β-aggregation propensities measured from trimeric simulations is however not possible. Although by definition the order parameters do not depend on the number of molecules (n) in the simulation box, when this number is rather small a n-dependent "background" order[3] is detected[4]. β-aggregation propensities obtained from hexameric and trimeric trajectories are therefore not directly comparable. Nonetheless, if one considers only triplets of peptides from each hexameric snapshot, the order information becomes homogeneous and different oligomeric systems can be compared. To guarantee a fair comparison, only triplets of non-isolated peptides, i.e., triplets where each chain forms at least a C_α contact ($C_\alpha - C_\alpha < 5.5$ Å) with another chain, were considered. Hexameric and trimeric amyloidogenicity profiles are shown in Fig. S2 (blue and red circles, respectively). Remarkably, the correlation between the two is quantitative and the grey bars, which show β-aggregation differences ($\Delta\beta$) along the profile, are rather small (below 0.05) almost everywhere. For both systems three *hot-spots* separated by non aggregation-prone regions have been identified. Interestingly, the C-terminal *hot-spot* in the trimeric profile (red spots) was not originally observed (see Fig. 1). Here, the third aggregation-prone region originates from the restriction of the order analysis to non-isolated peptide triplets and gives further evidence of the role played by the condensation equilibrium in amyloid aggregation (see previous section). Free energy profiles along R_g, used to monitor the condensation propensity of the hexameric system, were also measured (data not shown); in agreement with previous findings, the C-terminal stretches

(from $A\beta_{27-33}$ to $A\beta_{35-41}$) showed little condensation propensity.

The structural properties of the aggregates sampled by the MD simulations were also analyzed. By cluster analysis, the trajectories of the stretches located in the *hot-spots* revealed the existence of fully ordered states. The majority of these low-energy states were structurally diverse six-stranded β-sheets characterized by different register alignment, twist and polarity content. Although a statistically relevant preference for in-register parallel arrangements was observed, no hexameric system exhibited a single dominating free energy minimum. This finding is consistent with the very recent discovery that the well-known amyloid-fibril polymorphism results from significant variations in the molecular structure at the protofilament level[5]. The observed "fibril-like" aggregates are fully consistent with the ordered arrangements sampled along the trimeric trajectories and can be thought as their logic extension. The emergence of multiple and competitive low free-energy states suggests, on one hand, that the critical nucleus size has not been reached yet and, on the other hand, that kinetically trapped states might be involved in the structural evolution of ordered oligomers[6].

4. Effect of the segment length on β-aggregation

The robustness of the simulation results with respect to the length of the segments that are identified along the polypeptide sequence was also investigated. The $A\beta_{42}$ primary structure was decomposed into overlapping hendecamers (Table S2) and the trimeric systems of individual segments were studied by implicit solvent MD simulations. The increasing complexity arising from the presence of longer peptide chains resulted in insufficient sampling of the conformational space by constant temperature MD. In fact, a few preliminary runs at 330 K produced trajectories where the system remained trapped in a single free-energy minimum for the whole length of the simulation ($\sim 2\mu$s, data not shown). Thus, replica exchange MD (REMD)[4] was preferred to investigate the aggregation process of 11-residue stretches. Profiles of β-aggregation propensity as a function of temperature are shown in Fig. S3. At elevated temperature, the profiles converge to 0.5, which indicates that no orientational order is present and different stretches are indistinguishable. Close to the physiological temperature, a strong heterogeneity becomes apparent: (i) the stretches centered at residues 14 and 16 show a steep sigmoidal profile which means large β-aggregation propensity over a

wide range of temperatures; (ii) stretches 6, 12, 18 and 20 display a maximum indicating the existence of an equilibrium that disfavors ordered states at low temperature; (iii) stretches 8 and 22 show monotonically growing profiles which resemble the behavior of highly amyloidogenic sequences, though shifted to lower temperatures; and (iv) the remaining eight stretches (10, 24, 26, 28, 30, 32, 34 and 36) do not show any β-aggregation propensity over the range investigated. Whereas both monotonically growing and non-amyloidogenic profiles have been already observed for 7-residue peptides[4], trends characterized by a maximum are detected here for the first time. Secondary structure analysis (data not shown) revealed that the competition between self (intramolecular) and cross (intermolecular) interactions is responsible for such behavior. By lowering the temperature the self interactions (e.g., α-helical and loop conformations) of certain segments are favored at the expenses of the extended conformations and the β-aggregation propensity drops.

The values of β-aggregation propensity from implicit solvent REMD trajectories were used to draw the $A\beta_{42}$ amyloidodgenicity profile at 330 K. However, to compare the results obtained from different segment decompositions (i.e., 7- and 11-residue peptide segments) the profile was re-derived by considering 7-residue substretches on the 11-residue segments. For each hendecapeptide three substretches were identified (e.g., $\mathtt{D}_1\mathtt{AEFRHD}_7$, $\mathtt{E}_3\mathtt{FRHDSG}_9$ and $\mathtt{R}_5\mathtt{HDSGYE}_{11}$ from $\mathtt{D}_1\mathtt{AEFRHDSGYE}_{11}$; $\mathtt{E}_3\mathtt{FRHDSG}_9$, $\mathtt{R}_5\mathtt{HDSGYE}_{11}$ and $\mathtt{D}_7\mathtt{SGYEVH}_{13}$ from $\mathtt{E}_3\mathtt{FRHDSGYEVH}_{13}$; and so on) and the time series of the nematic order parameter was computed for each of them along the trajectory. β-Aggregation profiles obtained from 7- and 11-residue segment simulations are comparable (Fig. S4, top); the trends are very similar and display a maximum in the region 10-22 and two minima at $\mathtt{S}_8\mathtt{G}_9$ and the C-terminus. Due to the errors inherent to the implicit solvation model, the N-terminus is more aggregation-prone than the C-terminus in both cases. Similar conclusions can be drawn from the comparison of the β-sheet structure profiles (Fig. S4, bottom). Again, maxima and minima are similarly located along the $A\beta_{42}$ sequence. The analysis shows that the segment length does not affect the simulation results.

5. Structural interpretation of the $A\beta_{42}$ aggregation profile

The $A\beta_{42}$ β-aggregation profile from simulations of three 11-residue peptides (Table S2) shows a high propensity region encompassing residues 14-22 (Fig. 2, top). Interestingly,

β-aggregation propensities indicate the following ranking $14 \approx 16 \gg 18 > 20 \gg 22$. To structurally characterize the amyloidogenic trend in this central *hot-spot*, a secondary structure analysis of the conformations saved along the trajectories was performed. Secondary structure histograms plotted along the sequence are shown in Fig. S5. A α-helical/β-strand equilibrium, consistent with NMR studies that highlighted a trend to helical structures for the segment 16-24 in aqueous solution[7], accounts for the shape of the profile. On the left edge of the central *hot-spot* (stretch 12), α-helical conformations are dominating and very low β-aggregation propensity is observed. In stretches 14 and 16 the β-strand content increases dramatically, the α-helical propensity vanishes and the sequences are highly amyloidogenic. In stretches 18 and 20 the α/β equilibrium restores and although β-strand conformations are still preferred at this stage, α-helical structures become more stable. Finally, in stretch 22 a turn-like motif ($\mathtt{G}_{25}\mathtt{S}_{26}$) shows up and a second large drop in the profile occurs. Hence, β-aggregation propensity correlates with β-strand content, anticorrelates with α-helical content and seems very sensitive to β-turn or bend propensity. The simulation results indicate that both α-helical/β-strand and β-turn/β-strand equilibria modulate the aggregation properties of this region.

6. Structure stability from explicit water runs

For a subset of four $A\beta_{42}$ 7-residue stretches ($A\beta_{1-7}$, $A\beta_{5-11}$, $A\beta_{13-19}$, and $A\beta_{29-35}$) explicit solvent MD simulations were started from in-register parallel β-sheet conformations observed in implicit solvent runs to evaluate their structural stability. For each peptide system, 50-ns explicit water production runs were performed and the time series of $\overline{P_2}$ were computed along the trajectories. Average values taken at time intervals of increasing length (0.002, 1, 2, 5, 10, 20, 30, and 50 ns) are shown in Fig. S6. The quick drop of the average $\overline{P_2}$ in the case of the $A\beta_{1-7}$ and $A\beta_{5-11}$ segments indicates marginal structural stability of the in-register parallel β-sheet arrangements. On the contrary, segments $A\beta_{13-19}$ and $A\beta_{29-35}$ show rather stable β-sheet conformations in agreement with experimental observations[8–10].

---

[1] P. Ferrara, A. Caflisch, Native topology or specific interactions: What is more important for peptide folding?, J. Mol. Biol. 306 (2001) 837–850.

[2] A. Cavalli, U. Haberthür, E. Paci, A. Caflisch, Fast protein folding on downhill energy landscape, Protein Science 12 (2003) 1801–1803.

[3] T. P. Doerr, D. Herman, H. Mathur, P. L. Taylor, Randomness in nanoscopic liquid-crystal droplets - how small is small?, Europhys. Lett. 59 (3) (2002) 398–402.

[4] M. Cecchini, F. Rao, M. Seeber, A. Caflisch, Replica exchange molecular dynamics simulations of amyloid peptide aggregation, J. Chem. Phys. 121 (21) (2004) 10748–10756.

[5] A. T. Petkova, R. D. Leapman, Z. Guo, Y. W., M. P. Mattson, R. Tycko, Self-propagating, molecular level polymorphism in Alzheimer's β-amyloid fibrils, Science 307 (2005) 262–265.

[6] W. Hwang, Z. Shuguang, R. D. Kamm, M. Karplus, Kinetic control of dimer structure formation in amyloid fibrillogenesis, Proc. Natl. Acad. Sci. USA. 101 (35) (2004) 12916–12921.

[7] R. Riek, P. Güntert, H. Döbeli, B. Wipf, K. Wüthrich, NMR studies in aqueous solution fail to identify significant conformational differences between the monomeric forms of two Alzheimer peptides with widely different plaque-competence, Aβ(1-40)^{ox} and Aβ(1-42)^{ox}, Eur.J.Biochem. 268 (2001) 5930–5936.

[8] L. O. Tjernberg, J. Näslund, F. Lindqvist, J. Johansson, A. R. Karlstrom, J. Thyberg, L. Terenius, C. Nordstedt, Arrest of β-amyloid fibril formation by a pentapeptide ligand., J. Biol. Chem. 271 (1996) 8545–8548.

[9] A. T. Petkova, Y. Ishii, J. J. Balbach, O. N. Antzutkin, R. D. Leapman, F. Delaglio, R. Tycko, A structural model for Alzheimer's β-amyloid fibrils based on experimental constraints from solid state NMR, Proc. Natl. Acad. Sci. USA. 99 (26) (2002) 16742–16747.

[10] R. Liu, C. McAllister, Y. Lyubchenko, M. R. Sierks, Residues 17-20 and 30-35 of β-amyloid play crucial roles in aggregation, J. Neurosci. Res. 75 (2003) 162–171.

TABLE S1: $A\beta_{42}$: 7-Residue stretches simulations of four disease-related and one non-pathological variants

Segment	Variant	Peptide Sequence	Central Residue	3 peptides (μs) 310 K	330 K
$A\beta_{15-21}$	A21G	DAEFRHDSGYEVHH**QKLVFFG**EDVGSNKGAIIGLMVGGVVIA	18	3×1.0	
$A\beta_{17-23}$	A21G	DAEFRHDSGYEVHHQK**LVFFGED**VGSNKGAIIGLMVGGVVIA	20	3×1.0	
$A\beta_{19-25}$	A21G	DAEFRHDSGYEVHHQKLV**FFGEDVG**SNKGAIIGLMVGGVVIA	22	3×1.0	
$A\beta_{21-27}$	A21G	DAEFRHDSGYEVHHQKLVFF**GEDVGSN**KGAIIGLMVGGVVIA	24	3×1.0	
$A\beta_{17-23}$	E22G	DAEFRHDSGYEVHHQK**LVFFAGD**VGSNKGAIIGLMVGGVVIA	20	3×1.0	
$A\beta_{19-25}$	E22G	DAEFRHDSGYEVHHQKLV**FFAGDVG**SNKGAIIGLMVGGVVIA	22	3×1.0	
$A\beta_{21-27}$	E22G	DAEFRHDSGYEVHHQKLVFF**AGDVGSN**KGAIIGLMVGGVVIA	24	3×1.0	
$A\beta_{17-23}$	E22K	DAEFRHDSGYEVHHQK**LVFFAKD**VGSNKGAIIGLMVGGVVIA	20	3×1.0	
$A\beta_{19-25}$	E22K	DAEFRHDSGYEVHHQKLV**FFAKDVG**SNKGAIIGLMVGGVVIA	22	3×1.0	
$A\beta_{21-27}$	E22K	DAEFRHDSGYEVHHQKLVFF**AKDVGSN**KGAIIGLMVGGVVIA	24	3×1.0	
$A\beta_{17-23}$	E22Q	DAEFRHDSGYEVHHQK**LVFFAQD**VGSNKGAIIGLMVGGVVIA	20	3×1.0	
$A\beta_{19-25}$	E22Q	DAEFRHDSGYEVHHQKLV**FFAQDVG**SNKGAIIGLMVGGVVIA	22	3×1.0	
$A\beta_{21-27}$	E22Q	DAEFRHDSGYEVHHQKLVFF**AQDVGSN**KGAIIGLMVGGVVIA	24	3×1.0	
$A\beta_{13-19}$	F19S	DAEFRHDSGYEV**HHQKLVS**FAEDVGSNKGAIIGLMVGGVVIA	16		1×1.9
$A\beta_{15-21}$	F19S	DAEFRHDSGYEVHH**QKLVSFA**EDVGSNKGAIIGLMVGGVVIA	18		1×1.7
$A\beta_{17-23}$	F19S	DAEFRHDSGYEVHHQK**LVSFAED**VGSNKGAIIGLMVGGVVIA	20		1×1.3
$A\beta_{19-25}$	F19S	DAEFRHDSGYEVHHQKLV**SFAEDVG**SNKGAIIGLMVGGVVIA	22		1×1.3

TABLE S2: $A\beta_{42}$: 11-Residue stretches simulations

Segment	Peptide Sequence	Central Residue	3 peptides (μs) CTMD 330 K	3 peptides (μs) REMD 294-413 K
$A\beta_{1-11}$	**DAEFRHDSGYE**VHHQKLVFFAEDVGSNKGAIIGLMVGGVVIA	6	1×1.5	10×1.0
$A\beta_{3-13}$	DA**EFRHDSGYEVH**HQKLVFFAEDVGSNKGAIIGLMVGGVVIA	8	1×1.4	10×1.0
$A\beta_{5-15}$	DAEF**RHDSGYEVHHQ**KLVFFAEDVGSNKGAIIGLMVGGVVIA	10	1×1.3	10×1.0
$A\beta_{7-17}$	DAEFRH**DSGYEVHHQKL**VFFAEDVGSNKGAIIGLMVGGVVIA	12	1×1.5	10×1.0
$A\beta_{9-19}$	DAEFRHDS**GYEVHHQKLVF**FAEDVGSNKGAIIGLMVGGVVIA	14	1×1.5	10×1.0
$A\beta_{11-21}$	DAEFRHDSGY**EVHHQKLVFFA**EDVGSNKGAIIGLMVGGVVIA	16	1×1.5	10×1.0
$A\beta_{13-23}$	DAEFRHDSGYEV**HHQKLVFFAED**VGSNKGAIIGLMVGGVVIA	18	1×1.5	10×1.0
$A\beta_{15-25}$	DAEFRHDSGYEVHH**QKLVFFAEDVG**SNKGAIIGLMVGGVVIA	20	1×1.7	10×1.0
$A\beta_{17-27}$	DAEFRHDSGYEVHHQK**LVFFAEDVGSN**KGAIIGLMVGGVVIA	22	1×1.8	10×1.0
$A\beta_{19-29}$	DAEFRHDSGYEVHHQKLV**FFAEDVGSNKG**AIIGLMVGGVVIA	24	1×1.7	10×1.0
$A\beta_{21-31}$	DAEFRHDSGYEVHHQKLVFF**AEDVGSNKGAI**IGLMVGGVVIA	26	1×1.9	10×1.0
$A\beta_{23-33}$	DAEFRHDSGYEVHHQKLVFFAE**DVGSNKGAIIG**LMVGGVVIA	28	1×2.0	10×1.0
$A\beta_{25-35}$	DAEFRHDSGYEVHHQKLVFFAEDV**GSNKGAIIGLM**VGGVVIA	30	1×2.0	10×1.0
$A\beta_{27-37}$	DAEFRHDSGYEVHHQKLVFFAEDVGS**NKGAIIGLMVG**GVVIA	32	1×2.1	10×1.0
$A\beta_{29-39}$	DAEFRHDSGYEVHHQKLVFFAEDVGSNK**GAIIGLMVGGV**VIA	34	1×3.0	10×1.0
$A\beta_{31-41}$	DAEFRHDSGYEVHHQKLVFFAEDVGSNKGA**IIGLMVGGVVI**A	36	1×2.3	10×1.0

TABLE S3: hIAPP_{1-37}: 7-Residue stretches simulations

Segment	Peptide Sequence	Central Residue	3 peptides (μs) 330 K
hIAPP_{1-7}	**KCNTATC**ATQRLANFLVHSSNNFGAILSSTNVGSNTY	4	1×1.1
hIAPP_{3-9}	KC**NTATCAT**QRLANFLVHSSNNFGAILSSTNVGSNTY	6	1×1.1
hIAPP_{5-11}	KCNT**ATCATQR**LANFLVHSSNNFGAILSSTNVGSNTY	8	1×1.1
hIAPP_{7-13}	KCNTAT**CATQRLA**NFLVHSSNNFGAILSSTNVGSNTY	10	1×1.1
hIAPP_{9-15}	KCNTATCA**TQRLANF**LVHSSNNFGAILSSTNVGSNTY	12	1×1.1
hIAPP_{11-17}	KCNTATCATQ**RLANFLV**HSSNNFGAILSSTNVGSNTY	14	1×1.1
hIAPP_{13-19}	KCNTATCATQRL**ANFLVHS**SNNFGAILSSTNVGSNTY	16	1×1.1
hIAPP_{15-21}	KCNTATCATQRLAN**FLVHSSN**NFGAILSSTNVGSNTY	18	1×1.1
hIAPP_{17-23}	KCNTATCATQRLANFL**VHSSNNF**GAILSSTNVGSNTY	20	1×1.1
hIAPP_{19-25}	KCNTATCATQRLANFLVH**SSNNFGA**ILSSTNVGSNTY	22	1×1.1
hIAPP_{21-27}	KCNTATCATQRLANFLVHSS**NNFGAIL**SSTNVGSNTY	24	1×1.1
hIAPP_{23-29}	KCNTATCATQRLANFLVHSSNN**FGAILSS**TNVGSNTY	26	1×1.1
hIAPP_{25-31}	KCNTATCATQRLANFLVHSSNNFG**AILSSTN**VGSNTY	28	1×1.1
hIAPP_{27-33}	KCNTATCATQRLANFLVHSSNNFGAI**LSSTNVG**SNTY	30	1×1.1
hIAPP_{29-35}	KCNTATCATQRLANFLVHSSNNFGAILS**STNVGSN**TY	32	1×1.1
hIAPP_{31-37}	KCNTATCATQRLANFLVHSSNNFGAILSST**NVGSNTY**	34	1×1.1

TABLE S4: Ure2p$_{20-70}$: 7-Residue stretches simulations

Segment	Peptide Sequence	Central Residue	3 peptides (μs) 330 K
Ure2p$_{20-26}$	**NIGNRNS**NTTTDQSNINFEFSTGVNNNNNNNSSSNNNNVQNNNSGRNGSQN	23	1×1.8
Ure2p$_{22-28}$	NI**GNRNSNT**TTDQSNINFEFSTGVNNNNNNNSSSNNNNVQNNNSGRNGSQN	25	1×1.8
Ure2p$_{24-30}$	NIGN**RNSNTTT**DQSNINFEFSTGVNNNNNNNSSSNNNNVQNNNSGRNGSQN	27	1×1.7
Ure2p$_{26-32}$	NIGNRN**SNTTTDQ**SNINFEFSTGVNNNNNNNSSSNNNNVQNNNSGRNGSQN	29	1×2.0
Ure2p$_{28-34}$	NIGNRNSN**TTTDQSN**INFEFSTGVNNNNNNNSSSNNNNVQNNNSGRNGSQN	31	1×2.1
Ure2p$_{30-36}$	NIGNRNSNTT**TDQSNIN**FEFSTGVNNNNNNNSSSNNNNVQNNNSGRNGSQN	33	1×1.8
Ure2p$_{32-38}$	NIGNRNSNTTTD**QSNINFE**FSTGVNNNNNNNSSSNNNNVQNNNSGRNGSQN	35	1×1.7
Ure2p$_{34-40}$	NIGNRNSNTTTDQS**NINFEFS**TGVNNNNNNNSSSNNNNVQNNNSGRNGSQN	37	1×1.6
Ure2p$_{36-42}$	NIGNRNSNTTTDQSNI**NFEFSTG**VNNNNNNNSSSNNNNVQNNNSGRNGSQN	39	1×1.8
Ure2p$_{38-44}$	NIGNRNSNTTTDQSNINF**EFSTGVN**NNNNNNSSSNNNNVQNNNSGRNGSQN	41	1×2.0
Ure2p$_{40-46}$	NIGNRNSNTTTDQSNINFEF**STGVNNN**NNNNSSSNNNNVQNNNSGRNGSQN	43	1×2.3
Ure2p$_{42-48}$	NIGNRNSNTTTDQSNINFEFST**GVNNNNN**NNSSSNNNNVQNNNSGRNGSQN	45	1×1.8
Ure2p$_{44-50}$	NIGNRNSNTTTDQSNINFEFSTGV**NNNNNNN**SSSNNNNVQNNNSGRNGSQN	47	1×1.6
Ure2p$_{46-52}$	NIGNRNSNTTTDQSNINFEFSTGVNN**NNNNNSS**SNNNNVQNNNSGRNGSQN	49	1×1.8
Ure2p$_{48-54}$	NIGNRNSNTTTDQSNINFEFSTGVNNNN**NNNSSSN**NNNVQNNNSGRNGSQN	51	1×2.1
Ure2p$_{50-56}$	NIGNRNSNTTTDQSNINFEFSTGVNNNNNN**NSSSNNN**NVQNNNSGRNGSQN	53	1×2.1
Ure2p$_{52-58}$	NIGNRNSNTTTDQSNINFEFSTGVNNNNNNNS**SSNNNNV**QNNNSGRNGSQN	55	1×1.9
Ure2p$_{54-60}$	NIGNRNSNTTTDQSNINFEFSTGVNNNNNNNSSS**NNNNVQN**NNSGRNGSQN	57	1×1.7
Ure2p$_{56-62}$	NIGNRNSNTTTDQSNINFEFSTGVNNNNNNNSSSNN**NNVQNNN**SGRNGSQN	59	1×1.7
Ure2p$_{58-64}$	NIGNRNSNTTTDQSNINFEFSTGVNNNNNNNSSSNNNN**VQNNNSG**RNGSQN	61	1×1.7
Ure2p$_{60-66}$	NIGNRNSNTTTDQSNINFEFSTGVNNNNNNNSSSNNNNVQ**NNNSGRN**GSQN	63	1×1.1
Ure2p$_{62-68}$	NIGNRNSNTTTDQSNINFEFSTGVNNNNNNNSSSNNNNVQNN**NSGRNGS**QN	65	1×2.1
Ure2p$_{64-70}$	NIGNRNSNTTTDQSNINFEFSTGVNNNNNNNSSSNNNNVQNNNS**GRNGSQN**	67	1×2.0

TABLE S5: Ure2p$_{20-70}$: 7-Residue stretches simulations of single- and double-point N-to-S substitutions at positions 47, 48 and 49.

Segment	Variant	Peptide Sequence	Central Residue	3 peptides (μs) 330 K
Ure2p$_{44-50}$	N47S	NIGNRNSNTTTDQSNINFEFSTGV**NNNSNNN**SSSNNNNVQNNNSGRNGSQN	47	1×1.5
Ure2p$_{44-50}$	N48S	NIGNRNSNTTTDQSNINFEFSTGV**NNNNSNN**SSSNNNNVQNNNSGRNGSQN	47	1×1.5
Ure2p$_{44-50}$	N49S	NIGNRNSNTTTDQSNINFEFSTGV**NNNNNSN**SSSNNNNVQNNNSGRNGSQN	47	1×0.9
Ure2p$_{44-50}$	N4748S	NIGNRNSNTTTDQSNINFEFSTGV**NNNSSNN**SSSNNNNVQNNNSGRNGSQN	47	1×1.0
Ure2p$_{44-50}$	N4749S	NIGNRNSNTTTDQSNINFEFSTGV**NNNSNSN**SSSNNNNVQNNNSGRNGSQN	47	1×1.0
Ure2p$_{44-50}$	N4849S	NIGNRNSNTTTDQSNINFEFSTGV**NNNNSSN**SSSNNNNVQNNNSGRNGSQN	47	1×1.0
Ure2p$_{42-48}$	N4748S	NIGNRNSNTTTDQSNINFEFST**GVNNNSS**NNSSSNNNNVQNNNSGRNGSQN	45	1×1.4
Ure2p$_{44-50}$	N4748S	NIGNRNSNTTTDQSNINFEFSTGV**NNNSSNN**SSSNNNNVQNNNSGRNGSQN	47	1×1.0
Ure2p$_{46-52}$	N4748S	NIGNRNSNTTTDQSNINFEFSTGVNN**NSSNNSS**SNNNNVQNNNSGRNGSQN	49	1×1.4
Ure2p$_{48-54}$	N4748S	NIGNRNSNTTTDQSNINFEFSTGVNNNN**SNNSSSN**NNNVQNNNSGRNGSQN	51	1×1.5

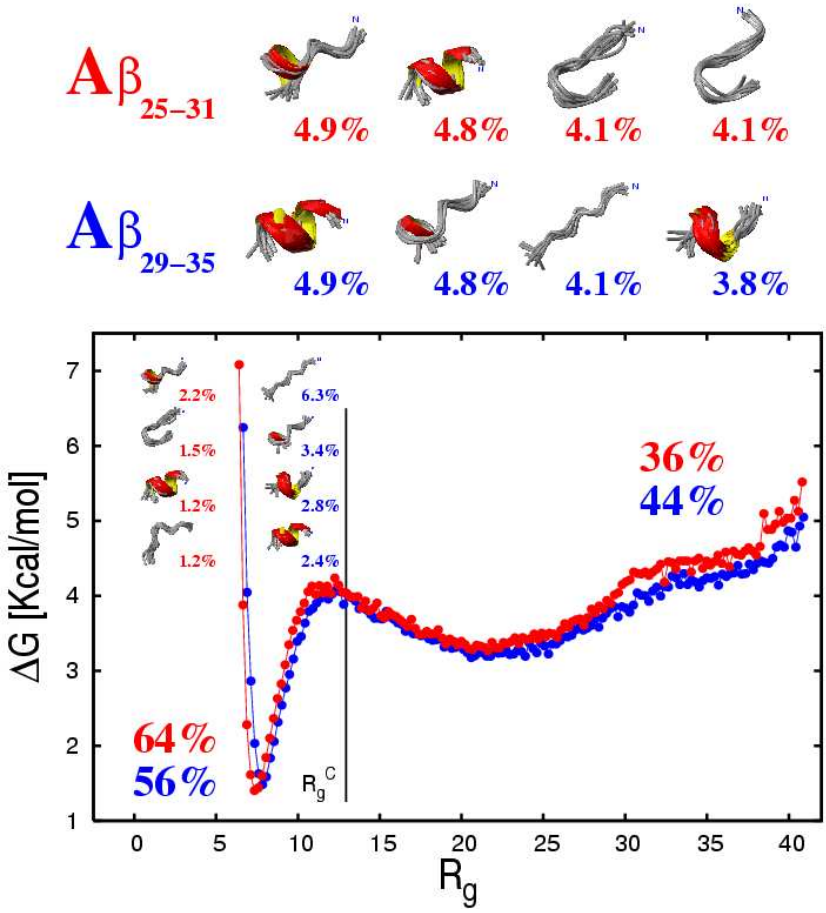


FIG. S1: Analysis of the non aggregation-prone 7-residue stretches $A\beta_{25-31}$ (red) and $A\beta_{29-35}$ (blue). (Top) Single-peptide cluster analysis on the whole trajectory. (Bottom, inset) Cluster analysis for fraction of conformations with $R_g < {R_g}^C$. ${R_g}^C$ corresponds to the lowest radius of gyration detected for conformations where all inter-peptide atomic distances are larger than the long-range interactions cutoffs (7.5 Å in this case). For both fractions (the whole and the condensed), the most populated clusters are represented with their statistical weight. (Bottom, main plot), free-energy projection along the radius of gyration of the trimeric system. Unlike the free-energy profiles on R_g, the single-peptide cluster analysis shows that the two trimeric systems are distinguishable and reveals the role played by the condensation equilibrium in β-aggregation.

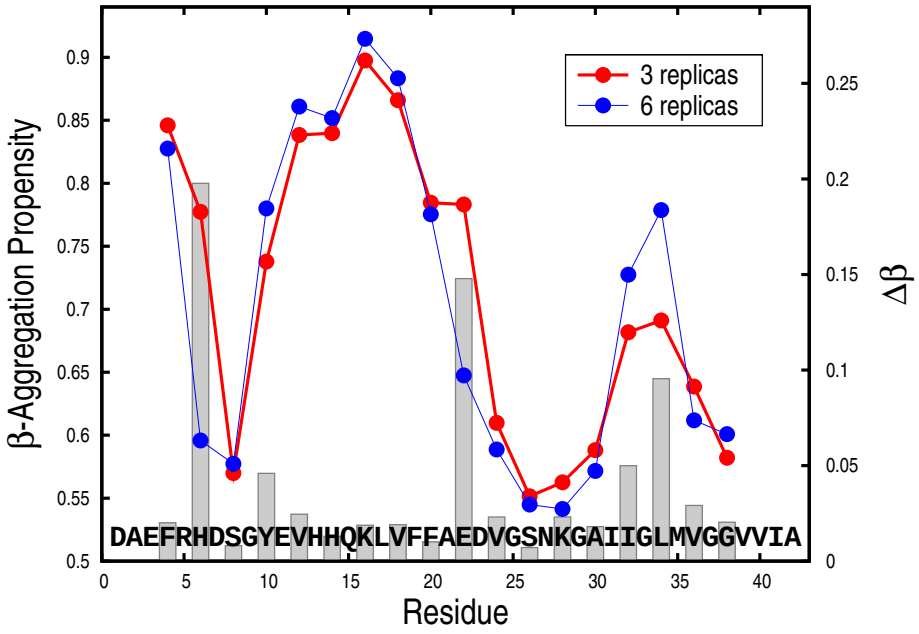


FIG. S2: Analysis of dependence on the size of the simulation system. Values of the β-aggregation propensity from 330 K constant temperature MD simulations of trimeric (red) and hexameric (blue) 7-residue peptide systems. To compare the different oligomeric systems, only triplets where any chain forms at least one C_α contact (C_α – C_α < 5.5 Å) with another chain were considered. β-Aggregation propensity differences ($\Delta\beta = \|\beta^{3rep} - \beta^{6rep}\|$) shown as gray bars with y-axis legend on the right highlight the largest discrepancies.

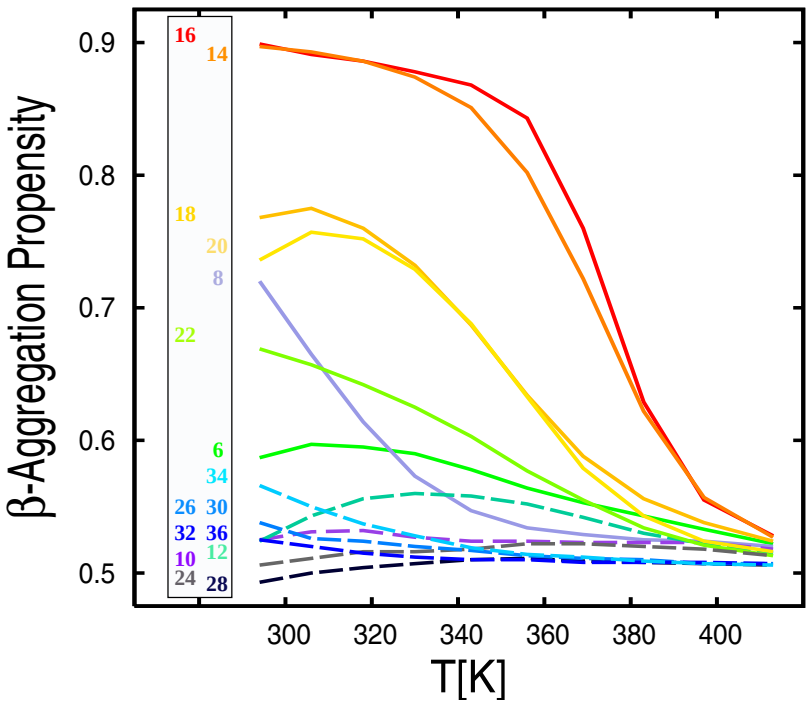


FIG. S3: $A\beta_{42}$: β-Aggregation propensity profiles as a function of temperature from REMD simulations of trimeric systems (11-residue peptide segments, Table S2 in Supplementary Material). Stretches with high β-aggregation propensity at physiological temperature values are highlighted by solid lines. The segment identification numbers, which are reported in the vertical panel on the left, correspond to the position of the central residue in the $A\beta_{42}$ full-length sequence.

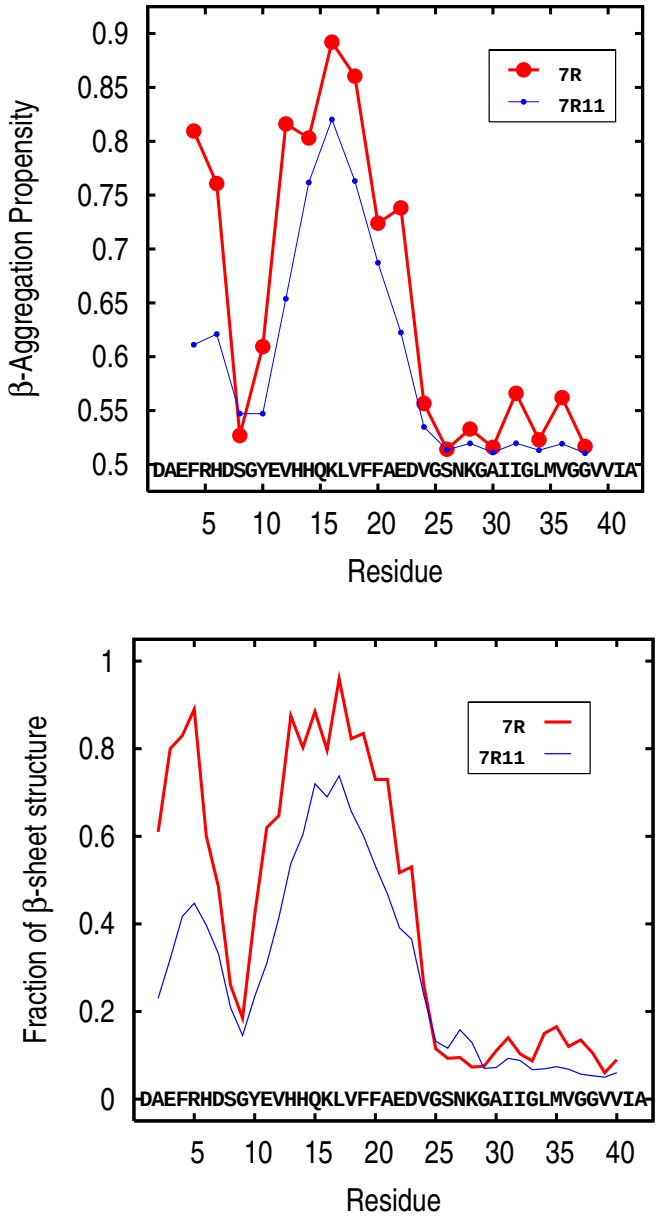


FIG. S4: Analysis of dependence on the length of the segments considered along the $A\beta_{42}$ sequence. β-Aggregation (top) and β-sheet structure (bottom) propensities have been extracted from 330 K MD trajectories of trimeric 7-residue (red) and 11-residue (blue) peptide systems. To allow for a direct comparison, β-aggregation and β-sheet structure propensities extracted from the 11-residue segment simulations have been computed by considering only 7-residue substretches (e.g., $\mathtt{D}_1\mathtt{AEFRHD}_7$, $\mathtt{E}_3\mathtt{FRHDSG}_9$ and $\mathtt{R}_5\mathtt{HDSGYE}_{11}$ from $\mathtt{D}_1\mathtt{AEFRHDSGYE}_{11}$).

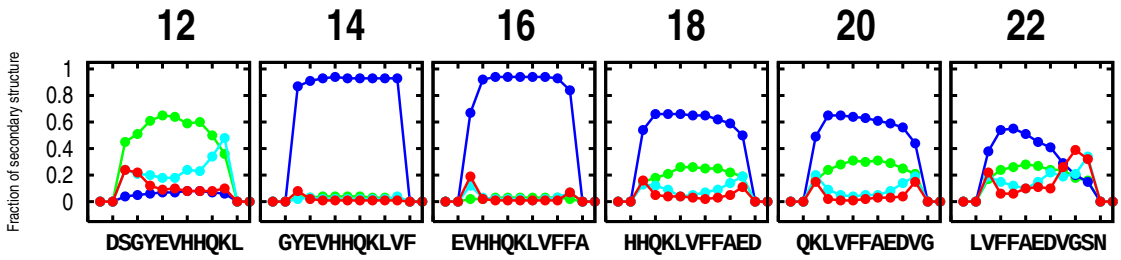


FIG. S5: Secondary structure histograms over REMD trajectory segments at 306 K for six $A\beta_{42}$ 11-residue central stretches. Green, blue, red and cyan dots correspond to α-helical, β-strand, β-turn or bend and random coil content, respectively. The histograms highlight an α-helical/β-strand competition in the central region of the $A\beta_{42}$ sequence.

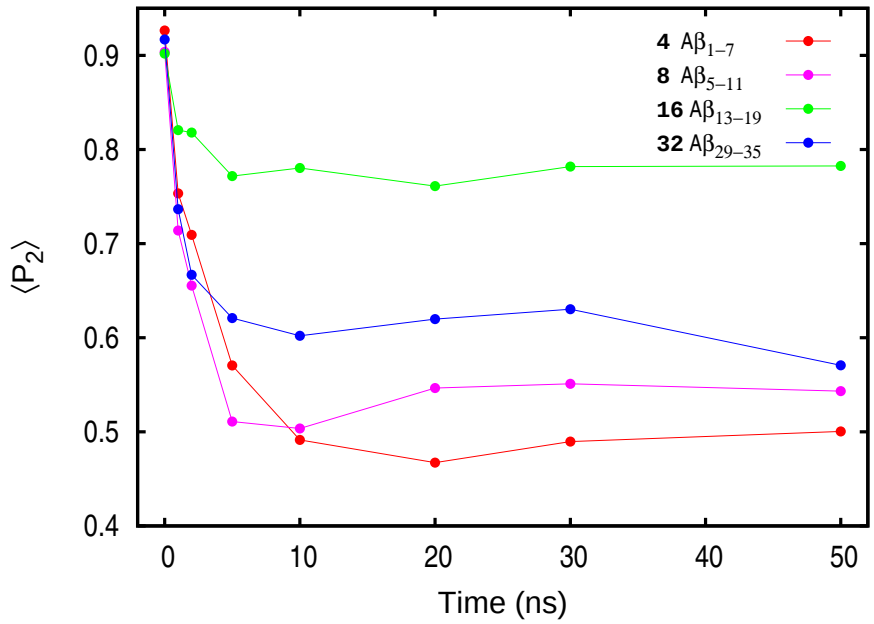


FIG. S6: Average $\overline{P_2}$ values computed from explicit solvent trajectories started from parallel β-sheet conformations observed in implicit solvent runs for four $A\beta_{42}$ 7-residue stretches: $A\beta_{1-7}$ (red), $A\beta_{5-11}$ (magenta), $A\beta_{13-19}$ (green), and $A\beta_{29-35}$ (blue). From left to right, the data points represent average values taken at time intervals of growing length: 0.002, 1, 2, 5, 10, 20, 30, and 50 ns.

Chapter 8

Conclusions

The relevance of computational studies in structural biology and their complementarity to *in vitro* biochemical experiments are today unquestionable. Although limited by non negligible statistical and systematic errors, computer simulations provide the ultimate detail concerning individual atom motion as a function of time, which are not accessible from a sample in a test tube. This detail can be used to describe the complete free-energy surface of individual proteins, which in the "new view" of protein folding represents the link between protein sequence and biological activity. In spite of continuous experimental progress this information is not available for any protein yet. Only the synergy between experimental and computational strategies can supply the description at the desired level of resolution. Computational approaches to simulate the behavior of model proteins at atomic resolution, called *in silico* experiments, must then be invoked. *In silico* experiments consist of three stages: (i) methodological development to solve particular problems; (ii) test-case application and comparison with available experimental data; and (iii) *blind* application for biological predictions. Following this paradigm, effective protocols can be designed to investigate difficult biological problems. In this thesis two *in silico* experiments have been presented: molecular docking for drug discovery and amyloid peptide aggregation. In the former, the in-house automatic approach for molecular docking has been improved, validated and successfully applied in a virtual screening project against β-secretase. In the latter, a novel molecular dynamics approach to investigate the aggregation properties of amyloid proteins has been developed. In agreement with experimental data, the strategy predicted the position dependence of β-aggregation propensity along the Alzheimer's peptide sequence.

The results reported in the thesis claim that *in silico* experiments for difficult problems, such as docking and amyloid peptide aggregation, can be successfully designed. An increasing role of computational studies in the understanding of complex biological processes is expected in the near future. Following the paradigm of the *in silico* experiment, molecular details

of abnormal processes associated to severe pathologies can be elucidated and efficient strategies for treating or even preventing such pathologies rationally developed.

List of Figures

Acknowledgments

I would like to thank Prof. Dr. Amedeo Caflisch for giving me the possibility to work in his research group and introducing me to the fascinating world of protein molecules. Furthermore, I would like to thank all the members of the group, former and current, for their help, advice and encouragement through all these years. Special thank to the "Wordom team", Francesco Rao and Dr. Michele Seeber, the "die-hard dockers", Dr. Shaheen Ahmed, Raffaele Curcio, Fabian Dey, Danzhi Huang, Peter Kolb, Dr. Stjepan Jelakovic and Dr. Nicolas Majeux, the "dynamic duo", Gian Gaetano Tartaglia and Riccardo Pellarin, the "professor" Enrico Guarnera, the "Matterhorn Admins", Christian Bolliger and Dr. Alexander Godknecht, and Dr. Rainer Böckmann, Dr. Andrea Cavalli, Dr. Jörg Gsponer, Dr. Urs Haberthür, Dr. Ronald Melki, Steffy Muff, Prof. Dr. Roger Nitsch, Dr. Emanuele Paci, Dr. Giovanni Settanni. Finally, I express my sincere gratitude to my family for continuous support and encouragement during my studies.

Curriculum vitae

Personal Details

Surname: CECCHINI
Name: Marco
Gender: Male
Date of birth: December 6, 1975
Place of birth: Bologna, Italy
Nationality: Italian

Education

10/2001 - now	Employed as PhD student at the University of Zurich, Prof. A. Caflisch – Biochemistry Department University of Zürich
9/1994 - 7/2000	Università degli Studi di Bologna (Italy) Master Degree ("Laurea") in Industrial Chemistry Final mark: 110/110 *summa cum laude* Theoretical thesis in Physical Chemistry: *"Modellazione e simulazione di mesogeni discotici chirali"* (Molecular modeling and simulation of chiral discotic mesogens) Prof. C. Zannoni – Dipartimento di Chimica Fisica ed Inorganica – Università degli Studi di Bologna
7/1994	High School Liceo Scientifico "A. Righi" – Bologna (Italy) Diploma of scientific high school Final mark: 60/60